Summer 8-10-2019

Identifying Risk Factors Related to Premature Birth Through Binary Logistic and Proportional Odds Ordinal Logistic Regression

Clayton Elwood

IDENTIFYING RISK FACTORS RELATED TO PREMATURE BIRTH THROUGH PROPORTIONAL ODDS ORDINAL LOGISTIC AND BINARY LOGISTIC REGRESSION

A Thesis

Submitted to the McAnulty College and Graduate School of Liberal Arts

Duquesne University

In partial fulfillment of the requirements for

the degree of Master of Science

By

Clayton Elwood

August 2019

IDENTIFYING RISK FACTORS RELATED TO PREMATURE BIRTH THROUGH PROPORTIONAL ODDS ORDINAL LOGISTIC AND BINARY LOGISTIC REGRESSION

By

Clayton Elwood

Approved June 24, 2019

Dr. Frank D'Amico Professor of Statistics (Committee Chair)	Dr. Stacy Levine Professor of Mathematics (Department Chair)
Dr. John Kern Associate Dean Associate Professor of Statistics (Committee Member)	Dr. Kristine Blair Dean, McAnulty College and Graduate School of Liberal Arts

ABSTRACT

IDENTIFYING RISK FACTORS RELATED TO PREMATURE BIRTH THROUGH BINARY LOGISTIC AND PROPORTIONAL ODDS ORDINAL LOGISTIC REGRESSION

By

Clayton Elwood

August 2019

Thesis supervised by Dr. Frank D'Amico

Premature birth has been identified as the single greatest cause of death worldwide in children under the age of five. This thesis will implement binary logistic regression and proportional odds ordinal logistic regression to predict different levels of premature birth and identify associated risk factors. The models will be built from the Center for Disease Control and Prevention's 2014 Vital Statistics Natality Birth Data containing nearly 4 million live births within the United States. Odds ratios and confidence intervals on risk factors were produced utilizing binary logistic regression.

ABSTRACT

OBJECTIVE

Identify and report risk factors associate with births prior to 37 weeks of gestation.

DESIGN

Observation Study

DATA SOURCES

Center for Disease Control and Prevention's 2014 Natality Public use file.

ELIGIBILITY CRITERIA FOR STUDY

First time mothers with no reported prior pregnancies. Mothers must have attended at least one recorded prenatal visit prior to delivery. Singleton births only. Case-wise deletion of missing values, dependent variable GESTREC10.

RESULTS

Mothers that attended six or fewer prenatal visits depending on the trimester of first prenatal care showed elevated odds of premature birth. Mothers starting care in the first trimester attending 1-6 prenatal visits have odds of premature birth 10.24 (9.93, 10.57) times greater compared to mothers with 11-16 prenatal visits. Mothers starting care in the second trimester attending 1-6 prenatal visits have odds of premature birth 3.54 (3.38, 3.71) times greater compared to mothers with 9-10 prenatal visits.

CONCLUSIONS

Mothers that attend fewer than the standard practice prenatal care plan may be at higher risk for premature birth.

ACKNOWLEDGEMENT

A special thank you to Dr. Frank D'Amico and Dr. John Kern for their remarkable guidance and mentorship during the completion of this thesis and the master's program. I would also like to thank Dr. James Schreiber from the School of Nursing for sharing his experience and knowledge of this important subject matter. Finally, a sincere thank you to my family, friends, and co-workers for their unwavering support.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1 – Introduction

1.1 Background

According to a recent study published in *The Lancet*, premature birth is the single greatest cause of death worldwide in babies and children under the age of 5.[1] Every year 1.09 million children die from complications linked to being delivered before 37 weeks of pregnancy.[1] In 2016, approximately 9.84 percent of all births in the United States were premature.[2] The most severe cases of premature birth, typically defined with a gestation period less than 28 weeks, pose the most long-term health risks and lowest survival rates. Identifying the factors that contribute to premature birth is a critical area of research because premature birth has also been shown to cause lifelong consequences for both the child and family. These issues include problems related to physical development, learning, communicating with others, getting along with others, and taking care of oneself. There are long-term disabilities linked to premature birth including behavior problems, Attention Deficit Hyperactivity Disorder (ADHD), anxiety, and neurological disorders (including Cerebral Palsy and Autism).[3]

This thesis utilizes the Center for Disease Control and Prevention's 2014 Natality Public use dataset, containing approximately 4 million birth records. The goal of this thesis is to explore different approaches to prediction and classification of premature birth with this dataset. All statistical methods used are detailed in Chapter 2. The typical data preparation and univariate analysis associated with studies of this type is documented in Chapter 3. In Chapter 4, the data is modeled using both Binary and Ordinal Logistic Regression. Findings from the overall analysis, including multivariate binary logistic regression odds ratios for all found risk factors, are reported and discussed in Chapter 5. Summary thesis remarks and future study are found in Chapter 6.

1.2 Risk Factors from Prior Research

Direct causes of premature birth remain unknown. A consensus of clinical research in the field has compiled a set of factors that pose elevated risk of premature birth. The Mayo Clinic has identified the following risk factors for premature birth:[4]

- Having a previous premature birth
- Pregnancy with twins, triplets, or other multiples
- An interval of less than six months between pregnancies
- Conceiving through in vitro fertilization
- Problems with the uterus, cervix or placenta
- Smoking cigarettes or using illicit drugs
- Some infections, particularly of the amniotic fluid and lower genital tract
- Some chronic conditions, such as high blood pressure and diabetes
- Being underweight or overweight before pregnancy
- Stressful life events, such as the death of a loved one or domestic violence
- Multiple miscarriages or abortions
- Physical injury or trauma

It is important to note that these factors only elevate the risk of premature delivery. The majority of mothers with identified risk factors deliver full-term and many mothers with no risk factors deliver prematurely.

Chapter 2 - Statistical Methods

The model form used in any statistical analysis is determined by the datatypes and level of measurement of the dependent variable (*Y*, response, or outcome) and independent variables (*X*, predictors, factors, or covariates). This chapter shows several models used for continuous outcome (Section 2.1), nominal outcome (Section 2.2), and ordinal outcome (Section 2.3) variables, while subsections of each focus upon changing the datatype of the independent variables. Some of these sections are presented for completeness.

2.1 Linear Regression Models

2.1.1 Continuous Outcome

When both independent (*X*) and dependent (*Y*) variables are continuous, the most basic method of prediction is fitting a line through a 2-dimensional graph, taking the linear form:

$$Y = b_0 + b_1 X + error. \quad \text{(Eq 1)}$$

Where b_0 is the y-intercept and b_1 represents the slope of the line that best fits *Y* to *X* with minimum error. Error is defined as the difference between the *Y* value of the prediction line and the observed *Y* value for a specific *X*. In 1806 French Mathematician, Adrien-Marie Legendre, developed the method of least squares in order to minimize the error term by minimizing the sum of squares:[5]

$$Sum\ of\ Squares\ error = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2. \quad \text{(Eq 2)}$$

From independently sampled random variables (X_i, Y_i) for $i = (1,2,..,n)$ where y_i is the i^{th} sample response and $b_0 + b_1 x_i$ is a line through this space at x_i, y_i with unknown parameters b_0, b_1. There exists a line with optimal parameter estimates $\hat{b}_0 \; and \; \hat{b}_1$ which minimize the sum of squares error. The methods to estimate these parameters include simulation, the method of least squares, the method of maximum likelihood, and computing the pseudoinverse (Moore–Penrose). The method of least squares and the method of maximum likelihood (under model assumptions specified below) give the following parameter estimates:

$$\hat{b}_1 = \frac{\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}\right]}{\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}\right]} .$$

Where $\bar{x}$ and $\bar{y}$ are the sample means. From $\hat{b}_1$ the intercept can be easily calculated:

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} .$$

Interpretation of the model's parameters are straightforward. For each one-unit increase in the independent variable X, the expected value of Y increases by $\hat{b}_1$. The assumptions required to use this model are a logical result of the model form, the error term (Eq 1) refers to the set of sum of square (Eq 2) errors:

- Linearity in the relationship between X and Y; since we are assuming a linear relationship for prediction
- The set (X, Y) are random samples from the population to ensure unbiased interpretation of the parameters
- The error term follows the normal distribution with $e \sim N(0, \sigma_e)$

- Homoscedasticity of the error term
- Errors are independent of one another, meaning they are random without significant correlation

These assumptions are required for inference on model parameters. The model (Eq 1) presented is commonly referred to as simple linear regression.

With additional independent continuous variables, the simple linear regression model is largely the same except that it attempts a linear fit through a higher dimensional space. While we could visually represent simple linear regression with a two dimensional graph, we cannot visually represent multilinear regression. However, the methods of estimation remain similar; determine the parameters that fit the data with minimum error. The multivariate model with k independent variables takes the form:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k + error.$$

As more and more predictors are added, the issue of overfitting the model will ultimately arise. Oxford Dictionary defines overfitting as "the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably." In multilinear regression, overfitting will occur by adding additional predictors in a model that effectively increases the dimensions of the solution space, alone explaining all the variance in the response variable. This phenomenon is called *the curse of dimensionality*, a term coined by Richard Bellman. In addition to overfitting caused by too many predictors, if any two or more predictors are highly correlated with one another, collinearity (multicollinearity when more than two) will occur. Collinearity and multicollinearity, unlike

overfitting, will not affect the predictive power of the model. However, parameter estimates become inaccurate to the true effect of individual predictors on the response. Small changes in the model can cause large changes in the estimates of parameters that correspond to highly correlated covariates. The model changes which parameter explains the same partition of the sum of squares error. In the extreme case of multicollinearity, two perfectly correlated independent variables, the resulting model will not converge and the optimal parameters do not exist.

When the datatype of a predictor variable is not continuous, analysis of variance (ANOVA) can be used to model a continuous response variable similar to what has been discussed thus far. ANOVA was developed by British Statistician, Sir Ronald Aylmer Fisher, and was published in *Statistical Methods and Scientific Inference* (1956). ANOVA relies on group means for each level of a nominal or ordinal independent variable α.

This model:

$$Y_{ij} = \mu + \alpha_i + error_{ij}$$

is a one-way ANOVA but is often referred to as an *Effects Model*. The term μ represents the mean of the entire sample. While Y_{ij} is the j^{th} observation in $(j = 1,2, \dots, n_i)$ of the i^{th} $(i = 1,2, \dots, k_j)$ effect (or treatment). Finally, α_i represents the difference between the ith group mean and the sample mean μ. Because of this, α can be viewed as the *Effect* of each treatment.

Parameter estimation can be performed mathematically by calculating the means: overall means, within-group means, and between-group means. From these the resulting sums of squares (SS) are calculated: SS total, SS treatment, and SS residuals. With these values the complete ANOVA table can be calculated. All of the assumptions from linear regression apply to the ANOVA model. Additional assumptions include:

- Variances in groups are equivalent (homogeneity of variance among treatment levels)
- Groups are independent

If the data includes both continuous and categorical independent variables, the simple linear regression and ANOVA can be combined. Blending ANOVA and simple linear regression together results in what is commonly referred to as an Analysis of Covariance (ANCOVA); additionally, multilinear regression yield MANCOVA models. ANCOVA models contain one continuous variable, often called the covariate, and one categorical variable. ANCOVA models are a possible improvement over ANOVA since you can include a continuous predictor. The model form:

$$Y = \mu + \alpha_i + b_1 X_1 + error_{adj}$$

splits total variance between the two datatypes; the categorical variable assumes a fixed shift between each level of the categorical variable like traditional ANOVA, while the continuous variable (covariate) helps explain within group variability, and in doing so will increase statistical power when this covariate is statistically significant. For example, when modeling a person's *body weight* as a function of *gender*. It would be more useful to add the covariate *height* into the model. The resulting model will show a stronger relationship between *body weight* and *gender* when *height* is accounted for by the model. The covariate is accounting for within group variation and allowing the remaining variability to be explained by group means (treatments). There are two model forms of ANCOVA differentiated by equal or unequal slopes. Interpretation of parameter estimates from an equal slope model can be interpreted exactly like ANOVA and Linear

Regression models discussed thus far, however unequal slope models cannot because of the interaction term. An interaction term is when two independent variables are crossed, typically by multiplication (e.g. *body weight* * *gender*), and the newly generated interaction term is also entered into the model. When this term is statistically significant the individual predictors and their interaction variable cannot be interpreted alone, each must be interpreted in relation to the value of its dependents.

Thus far, only changes to the independent variable datatype has been considered with a continuous response. However often times a researcher will seek to predict datatypes other than a continuous variable. These will be explored in the following sections and are the relevant methods used in this thesis. The discussion of continuous dependent models is for sake of completeness.

2.1.2 Nominal Outcome

A nominal variable is a type of categorical variable. Nominal type distinguishes between categories based on no assumed ordering of its levels, and at a minimum nominal data must contain two categories. In this minimal case, the datatype is often referred to as dichotomous or binary. Dichotomous variables, are a subset of nominal variables and assume either a value of zero or one which can represent categories (e.g. *Positive* = 1 and *Negative* = 0). Unlike continuous variables there is no measure of correlation between two nominal. Chi-square tests are typically used to understand the relationship between two (categorical) variables and in ways the test parallels correlation between two continuous variables. With a binary outcome and single binary explanatory variable, a 2x2 table can be formed from the sample with each combination to generate the observed cell counts. The Chi-square test statistic compares these observed cell counts to the expected counts (number of observations divided by number of cells); squaring each cell difference as a portion of expected cell counts and summing into the test statistic χ^2:

$$\chi^2 = \sum \frac{(observed\ count - Expected\ count)^2}{Expected\ count} .$$

The Chi-Square Test can theoretically be used with a sample of just one more than the number of cells (leaving a single degree of freedom for the test) but a general rule is to have no less than 5 observations per cell. The χ^2 test statistic follows the chi-square distribution that can lead to a p-value matching a two-sample z-test for a continuous outcome variable. Chi-square tests are the basis of logistic regression discussed in chapter 2.2.

2.1.3 Ordinal Outcome

Ordinal data is similar to nominal type; however, it assumes there is an underlying rank or order between each category. The relationship between the levels must be monotonic, however there is no fixed distance or scale between each level. An example of an ordinal variable is the Likert scale (disagree, neutral, agree) found in survey data.

2.1.3.1 Ordinal outcome – Nominal independent

There are several approaches for testing an ordinal outcome variable vs a nominal independent variable that are based on modified Chi-Square tests. Similarly, these models test expected cell counts vs the observed cell counts assuming some sort of trend is present as a result of the ordinal response variable.

2.1.3.2 Ordinal outcome – Ordinal independent

When one or both of the dependent and independent variables are ordinal, Log linear models can be used for testing relationships. Log Linear models are once again based on chi-squared tests and are non-predictive models. These models were developed by Alan Agresti in Categorical Data Analysis 1986. [6 (p314)]

2.1.3.3 Ordinal outcome – Continuous independent

When attempting to model expected cell counts from a continuous independent variable, a more sophisticated model is required that is based off a logistic regression model. Before looking at ordinal logistic regression in chapter 2.3, binary logistic regression will be explored in chapter 2.2.

2.2 Binary Logistic Regression

In the past sections we looked at methods of modeling continuous response variables which output the expected value of *Y* based on a set of independent predictors of varying datatypes. If the response variable is not continuous, let us assume it is binary. Then it would be preferable to model the probability that a given observation is either one or zero. Since the output of a simple linear model can take any value in (-∞,∞) using the linear predictor model $b_0 + b_1X$, a model that outputs a probability would need to be bounded to the interval [0, 1], and in doing so could be interpreted as a probability.

Let π represent $P(Y = 1)$. The odds that $Y = 1$ is the ratio of $\frac{\pi}{1-\pi}$. Taking the natural log of the odds yields the logistic transformation,

$$\log(odds) = \log\left(\frac{\pi}{1-\pi}\right). \qquad \text{(Eq 3)}$$

where the log*(odds)* can take any value from (-∞,∞) and return a unique value for $\log\left(\frac{\pi}{1-\pi}\right)$. Because of this, the logistic model can use the simple linear predictor $b_0 + b_1X$ to approximate log(*odds*):

$$\log\left(\frac{\pi}{1-\pi}\right) = b_0 + b_1 X. \qquad \text{(Eq 4)}$$

From this, we can solve (Eq 4) for π to get the probability form:

$$\pi = \frac{e^{b_0+b_1X}}{1+e^{b_0+b_1X}}. \qquad \text{(Eq 5)}$$

Parameter estimation of $\hat{b}_0, \hat{b}_1$ is approximated by maximum likelihood estimation; there is no equivalent least squares method. The distribution used is always determined by the datatype of the dependent variable. In this case, a binary dependent variable follows the Bernoulli distribution (a special case of the Binomial distribution). The likelihood for the π's when each y_i has the distribution $Bern(\pi_i)$ is:

$$\prod_{i=1}^{n} {\pi_i}^{y_i}(1-\pi_i)^{1-y_i}. \qquad \text{(Eq 6)}$$

By substituting in π (from Eq 5), the likelihood function in (Eq 6) can be expressed in terms of b_0, b_1:

$$L\,(b_0, b_1) = \prod_{i=1}^{n}\left[\frac{e^{b_0+b_1X}}{1+e^{b_0+b_1X}}\right]^{y_i}\left[\frac{e^{b_0+b_1X}}{1+e^{b_0+b_1X}}\right]^{1-y_i}.$$

Differentiating the natural log of the likelihood function ($\ln(L\,(b_0, b_1))$) with respect to b_0, b_1 to find the maximum would yield optimal parameter estimates $\hat{b}_0, \hat{b}_1$, however no closed formed

solution exists. Instead the parameters must be estimated through iterative algorithms such as Newton-Raphson; such algorithms are prevalent in modern statistical software.

The model discussed in this section is called the Logistic Regression Model, because it uses the logistic transformation (Eq 3). A major difference between Logistic Regression compared to the other models previously discussed is the lack of error term. The absence explains why we use the underlying Bernoulli distribution. Consider a simple logistic regression model with a binary response and a single nominal predictor with two levels. For each level of the predictor, each observation in that level is treated as a coin flip with odds equal to the parameter estimate for that level. The ramification of the missing error term means that all the tools available for model selection and diagnostics for linear regression (that centered entirely on the study of this error term) will not be available in logistic regression. Instead, the most common tools for diagnosing model fit in logistic regression surrounds the misclassification details in a confusion matrix. This thesis will utilize the following measures to test the predictive power of candidate models using a binary response (*Positive, Negative*):

- Misclassification Rate: the ratio of correctly classified observations (both positive and negative) over total observations
- Sensitivity: the ratio of correctly classified positive responses (true positive) over the total sample positives of the response
- Specificity: the ratio of correctly predicted negative responses (true negative) as a proportion of all negative responses in the sample
- Positive Predicted Value: the ratio of correctly classified positive responses
- Negative Predicted Value: the ratio of correctly classified negative responses

Sensitivity and specificity are extensively used in clinical research because it enables researchers to choose models that predict accurately in areas that are most important. For example, most initial disease test models seek to minimize specificity as the only priority with little regard for sensitivity or total misclassification rates, because it is preferable to have more false negative predictions than false positives. False positives lead to additional testing (and a possibly worried patient) while false negatives result in a sick patient believing they are healthy and not receiving the necessary treatment. Similarly, when modeling prematurity, model selection should focus on low specificity with high sensitivity with less regard for total accuracy. In addition, Lift and ROC curves can be used to visualize these measures (sensitivity and specificity) as a function of descending predicted values.

Interpretation of parameter estimates in logistic regression is not the same as linear models. For every one unit change in X, the model yields a one unit change in log*(odds)*. To overcome this limitation, the ratio of the odds at X and the odds at $X + 1$ are used instead, and are referred to as the odds ratio. An odds ratio can provide an analogous linear interpretation of logistic regression model parameter albeit multiplicative in nature. An odds ratio (for example using an odds ratio of 1.5 and treatment A versus B) is typically expressed by stating "the odds that Y=1 increase 1.5 times with treatment A versus treatment B."

2.3 Ordinal Logistic Regression – Proportional Odds model

The final model discussed in this chapter will be the Proportional Odds Ordinal Logistic Regression Model. Modeling with an ordinal dependent variable offers advantages over modeling each response level separately using binary logistic regression. While logistic regression on an ordinal dependent variable can take several forms, the most widely used is the proportional odds

model. With our Natality database it might be particularly useful to model a mother's risk for different severities of premature birth that pose even greater risk as detailed in Chapter 1.1. When modeling an ordinal response instead of a simple binary response, the output yields a cumulative probability for each level of the ordinal response. A mother's probability of having a *Normal birth week delivery*, *Premature* birth, or *Very Premature* birth would sum to 1. While separate binary logistic models would yield independent probabilities for each birth outcome, it is unlikely they will sum to one, there is also the possibility a mother would be classified into multiple response levels. The Proportional Odds Model solves these issues and offers a simplified interpretation of resulting odds ratios that will be detailed later in this section.

The Proportional Odds Model builds off the linear predictor used in binary logistic regression, however there are multiple intercepts for the ordinal response levels, while there remains a single coefficient estimate for each independent variable. The result of this setup is that the model assumes homogeneity of response for each level of the dependent variable for all independent variables over the prediction interval. The additional assumption on the Proportional Odds Model is crucial to test in order to construct a model that is capable of fitting the data and can only be used reliably over the range of already observed values. Multinomial Logistic Regression does not have the parallel slope assumption, and in some cases may provide an option for modeling an ordinal variable that violated this assumption as a nominal variable even though it is usually inadvisable to strip the important ordering information from ordinal data.

In order to model the ordinal response variable, the probability π from the binary logistic regression model in (Eq 5) is split into cumulative probabilities $\{\pi_1, ..., \pi_J\}$ for each j response level in J. [6(p 275)]. Then the probability that the ordinal response Y_i is less than or equal to a specific response level j is:

$$P(Y \leq j|x) = \pi_1(x) + \cdots + \pi_j(x)\,.$$

The cumulative logit link function describes the log(*odds*) of two cumulative probabilities measuring how likely an observation is in level *j* (or below) compared to above *j*:

$$\log\left[\frac{P(Y \leq j|x)}{P(Y > j|x)}\right] = \log\left(\frac{P(Y \leq j|x)}{1 - P(Y \leq j|x)}\right) = \log\left(\frac{\pi_1(x) + \cdots + \pi_j(x)}{\pi_{j+1}(x) + \cdots + \pi_J(x)}\right) \quad \text{(Eq 7)}$$

$$for\ j = 1, \ldots, J-1\,.$$

While the binary logistic model has a single logistic transformation link function, the Proportional Odds Model has $J-1$ such logits (Eq 7) for every level of the response. Note there are not J such levels, since the first level is derived as a function of the others. The logit link functions used in this thesis are detailed in Chapter 4.2. A Proportional Odds Model can use a linear predictor to approximate log(*odds*) just like binary logistic (Eq 4). However, each logit link will share the same parameter b_1 estimate for each predictor, but each response level j gets its own intercept α_j up to $J-1$:

$$P(Y \leq j|x) = \alpha_j + b_1 X \qquad j = 1, \ldots, J-1.$$

Optimal parameters occur at the maximum of the natural log of the likelihood function in (Eq 8). Parameter estimation resulting from Maximum Likelihood Estimation was originally presented by Walker and Duncan (1967) using a Fisher Scoring algorithm [6(p277)]:

$$L\left(\{\alpha_j\},\beta_1\right) = \prod_{i=1}^{n}\left[\prod_{j=1}^{J}\left[\frac{e^{\alpha_j+\beta_1 X}}{1+e^{\alpha_j+\beta_1 X}}\right]-\left[\frac{e^{\alpha_{j-1}+\beta_1 X}}{1+e^{\alpha_{j-1}+\beta_1 X}}\right]^{y_{ij}}\right]. \qquad \text{(Eq 8)}$$

With a model where J=4, for a fixed j the response curve is a logistic transformation curve for a binary response of outcomes $Y \leq$ j and $Y > j$ is defined by:

$$P(Y \leq j|x) = \frac{e^{(\alpha_j+b_1 X)}}{(1+e^{(\alpha_j+b_1 X)})} \qquad j = 1,\ldots,J-1. \qquad \text{(Eq 9)}$$

Depicted in Figure 1 are the response curves (Eq 9) displaced on the x-axis by the intercepts α_j, each exhibiting identical rate of change as a result of the shared parameter b_1. [6(p276)]

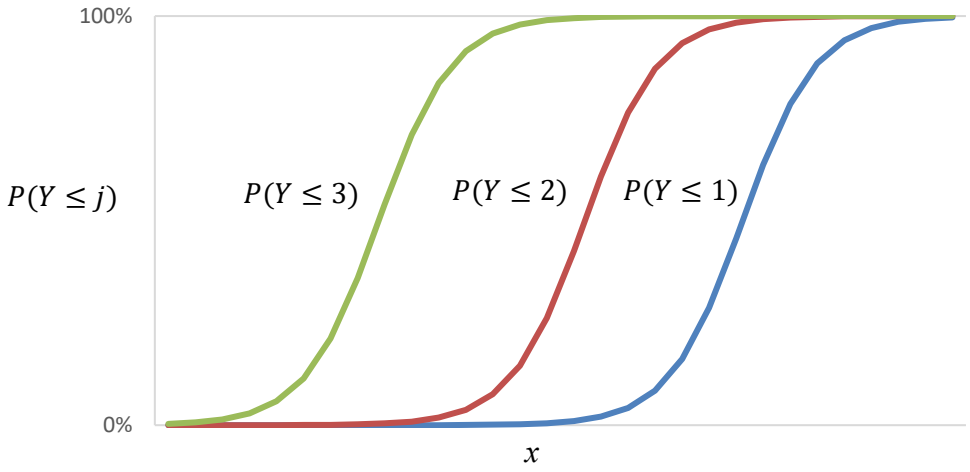


Figure 1 - Response Curves of Proportional Odds Model when J=4

Figure 2 depicts individual category probability distribution for each response level. The final probability curve j =1 is generated as a function of the other curves.

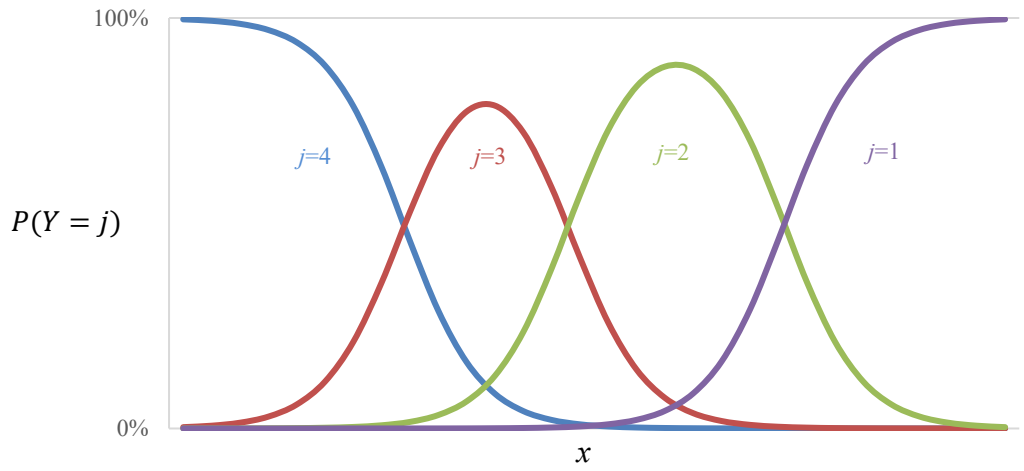


Figure 2 – Cumulative Probability Curves of Proportional Odds Model when J=4

Consequently, cumulative odds ratios exhibit a property that provides context for the name *Proportional Odds Model*. This means the odds of *Premature* birth compared to *Normal* birth for any given factor are proportional to the odds of *Very Premature* birth compared to *Premature* birth. As a result, for a one unit increase in X, the odds ratio is simply e^b for each factor. Odds ratio confidence intervals are derived from the Confidence Intervals around parameter estimates.

Chapter 3 - Data Preparation

3.1 *Natality Public Use File*

3.1.1 Properties

The Center for Disease Control and Prevention's *2014 Natality Public use file* will be used to build models for predicting different levels of premature birth and identifying risk factors associated with premature birth. The 2014 database contains 3,998,175 live births which took place within the United States to US citizens, legal residents, and non-resident aliens of the US in the calendar year 2014.[7] These births represent 100% of the birth certificates registered in all US states and the District of Columbia. Births within US territories (Puerto Rico, Virgin Islands, Guam, American Samoa, and Commonwealth of the Northern Marianas) are excluded from this file.[7] Births by US citizens born outside of the United States are also not included in the file. The Center for Disease Control and Prevention's (CDC) National Center for Health Statistics (NCHS) receives these data as electronic files, prepared from individual records processed by each registration area, through the Vital Statistics Cooperative Program.[8] It has been estimated by the CDC that 99% of all births within the United States are registered, and therefore this dataset can be considered representative of the total United States population of all births for this year. The dataset includes 241 variables, consisting of parental demographics, statistics measured during pregnancy, data collected during delivery, and finally detailed measures of the infant's health condition. In addition to raw data collected, NCHS imputes and recodes some of these measures using predefined methodologies detailed in each year's User Guide.[8] One such recode involves the dependent variable for this analysis, *GESTREC10*, which is a categorized version of birth week used to analyze prematurity in CDC research.[9]

3.1.2 Data Cleansing

This analysis will focus on a prediction of both binary and ordinal outcomes related to premature birth, thus the response variable will be a categorical version of (*GESTREC10*) birth week. Since the goal is to identify the factors that are related to premature birth, the first step is to reduce the dataset to only those variables that are useful towards this goal. Consequently, all post-birth and delivery variables were removed before any observation-reduction methods were executed. Case-wise deletion was used for any missing data in the 2014 file. The response variable (*GESTREC10)* contains 3,303 missing observations that were removed, leaving 3,994,872 remaining observations. According to the CDC, there is an extremely strong correlation between plural births (Table 1) and shorter gestations that influences the overall gestational age measure and "Analyzing births in singleton deliveries separately is important".[10] Table 1 below shows that multiple births is highly predictive of prematurity. Therefore, in this analysis we focus on singleton births.

Prevalence of Premature Birth by Plurality

Number of Births	% less than 37 weeks (Premature)	Frequency
1	9.62%	3,854,757
2	56.61%	135,571
3	93.58%	4,251
4	95.53%	246
5	100%	47
		3,994,872

Table 1 - Prevalence of Premature Birth by Plurality

As a result, the 140,115 plural births in 2014 were removed from the dataset reducing the observations down to 3,854,757. In order to ensure that the dataset is truly independent, only first

born births were kept in the final dataset. The decision to remove this large portion of the sample is based on the concern that a mother's prior pregnancies (whether they were pre-mature or not) may have an unaccounted for effect on the mother's behavior during her current pregnancy, thereby violating the assumption of independence and biasing findings. This decision resulted in the removal of 2,391,512 observations. Any mother that did not seek prenatal care before birth was also removed, resulting in the deletion of 15,533 observations. Lastly, all missing values in any independent variable resulted in the observation being removed. There were 133,467 observations where at least one independent variable had a missing or unknown value. A flowchart of each observation reduction step is depicted below in Figure 4, and shows a final dataset containing 1,314,245 observations.

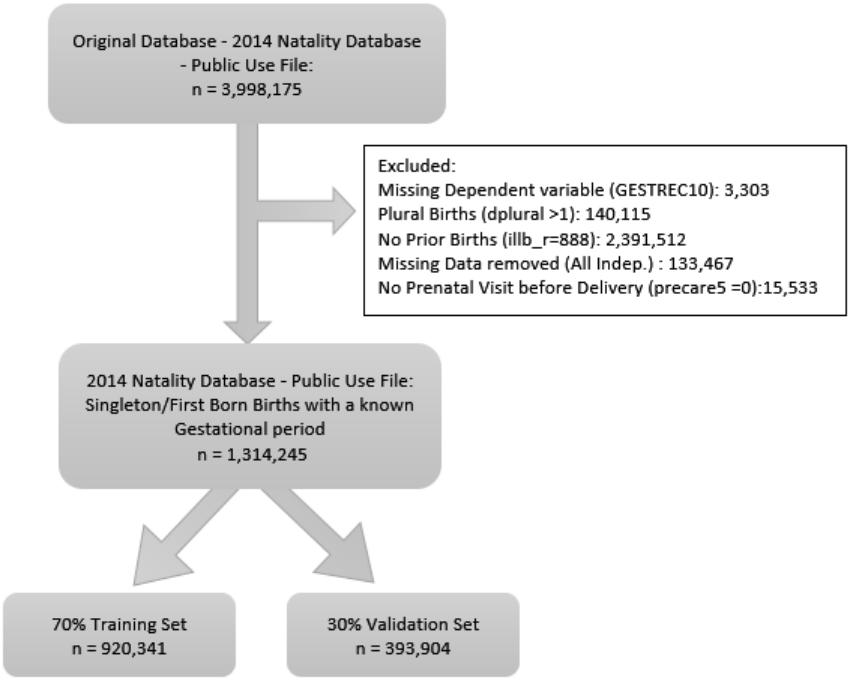


Figure 3 – Flowchart of Dataset Observation Reduction

While there has been a significant reduction (67%) of sample size, the underlying prevalence of *Premature* birth remains within 2% of the full dataset; 8.9% (n=1,314,245) compared to 10.34% (n=3,994,872) with a reported *GESTREC10 birth week*.

3.2 Outline of the Analysis

In section 4.1, we will start by dichotomizing the outcome variable birth week (*GESTREC10)* into two categories, either *Normal* birth or *Premature* birth. According to the World Health Organization (WHO), premature birth is any birth that occurs prior to the start of the 37 week of gestation.

For variable selection, the following method will be used to select the best variables that will make up the set of possible independent variables. First, standard univariate analysis will be used including Chi-Squared tests to rank each variable's relationship with our dichotomized response. Correlation between independent variables will be explored before narrowing down the best variables. Highly correlated variables will be evaluated individually choosing to keep those with the strongest relationship to the dependent variable, the fewest missing values, and/or the more meaningful variable for interpretation. For example, *Mother's Age* and *Father's Age* are both highly correlated with one another as well as being similarly predictive of *Premature* birth. However, the *Father's Age* variable had more missing observations and is less meaningful than *Mother's Age* for interpretation of the model, since not all pregnancies have a known father. For this reason, only *Mother's Age* was chosen in the final independent variable selection. Finally, a thorough investigation of the distributions for each variable will be performed during the final selection process. For continuous variables that are converted to nominal measures the binning methodology (if not performed by the CDC) will be based on a balance of even sample cell size and similar response level. If two nominal levels have similar prevalence of prematurity and are

not involved in a significant interaction term, these cells were combined for simplified reporting. Once all the independent variables have been selected, stepwise logistic techniques will be used to build a candidate model as detailed in Chapter 4.

In section 4.2, the response variable will be converted to an ordinal datatype with three levels representing *Normal* birth (greater than or equal to 37 weeks of gestation), *Premature* birth (between 32 and 37 weeks of gestation), and *Very Premature* birth (less than 32 weeks of gestation). A proportional odds model will be explored to determine the factors related to the 3-ordered categories of birth prematurity.

Chapter 4 - Model Implementation

4.1 Nominal Logistic Regression Model

Model construction utilized both SAS Institute's JMP Pro 13 software platform in conjunction with SAS BASE 9.4. The cleansed dataset containing 1,314,245 observations and 15 independent variables and was entered into the JMP stepwise logistic regression platform. All second degree interaction terms were included for each independent variable. The final dataset was split randomly into a 70% training set (920,341 observations) and a 30% validation set (393,904 observations) for testing model fit against observations not in the model. The final model's parameter estimates were confirmed with SAS Base 9.4. The next section will summarize common terms used in the JMP stepwise platform, several customization options available in the platform, and lastly the settings used for stepwise runs in this analysis.

4.1.2 Explanation of stepwise terms

Minimum Akaike Information Criterion (AIC) and related Corrected Akaike Information Criterion (AICc) were first developed by Hirotugu Akaike in 1973. Each model in the set of all possible models (based on every combination of the model's inputs) is assigned an AIC score using (Eq 10). The "best fitting" model for the dataset is the model with the lowest AIC score. Since the goal is to find the model form that leads to the lowest AIC score, we can again use maximum likelihood estimation on either AIC or AICc. AIC is defined as follows:

$$AIC = 2K - 2\log(L), \qquad \text{(Eq 10)}$$

where K is equal to the degrees of freedom of the candidate model, L represents the likelihood function at its maximum point for a given candidate model in the set of all possible models. AICc is an adaptation of AIC that simply adds a penalty for complex models proportional to their sample size, where n represents the sample size of the dataset in the following AICc equation:

$$AICc = AIC + \frac{2K(K+1)}{n-K-1} \cdot$$

The stepwise algorithm has several other options in addition to the measure used for maximization that can affect the output. These include step methodology and rules that define the treatment of effects. With minimum AICc, JMP only allows for forward and backwards steps, however JMP offers a combination of forward and backwards steps that is not based on AIC and will not be covered in this analysis. Forward selection stepping methodology begins with no independent variables in the model and chooses the single variable with the largest change in AICc. Forward selection continues until the stopping rule is reached, which is typically when the change in AICc is smaller than a given threshold. Backwards selection works in the opposite manner, starting with a fully saturated model and instead removing the least significant independent variable until no more insignificant terms can be removed from the model. The algorithm stops when the change in AICc is greater than a given threshold.

The Stepwise Platforms options provides the ability to customize how interaction terms and nominal variables with more than two levels are treated within the stepwise algorithm. The four settings detailed below can allow a model to be constructed in different ways.

- No Rules: Any term (or dummy variable created to represent a single or group of levels in a nominal independent variable) that is found significant can be added to the model regardless of hierarchy.
- Restrict: this method restricts any interaction term's entry into a model until it is precedent terms have been entered.
- Whole Effect: This technique prevents the creation of a model with missing effects, which means all levels of a nominal variable must be in the model or none at all.
- Combine Rule: The default JMP method performs two F-tests, first testing the group's significance probability for entry as a joint F test (P1). Next, a test of the significance after the precedent term has already been entered is performed (P2). The value used for entry criteria is maximum (P1, P2). In addition, the combine rule converts nominal and ordinal data into hierarchies using a tree structure similar to partitioning by maximizing the sum of squares between groups. The purpose of this is to maximize variance between the nominal groups. This comes at the possible cost of interpretation, as the algorithm may split, join, or exclude certain effect levels.[11]

Stepwise model evaluation will be based on the minimum Akaike Information Criterion (AIC) measure, using both forward and backwards selection, and both combined rules for discrete variables and the whole effect rule. The dependent variable for this model is the dichotomized *Premature* birth variable discussed in Chapter 3 section 2.

4.1.3 Descriptive Statistics and Definitions of Independent Variables

Table 2 summaries the descriptive statistic of the significant independent variables selected for entry into the stepwise platform as detailed in Chapter 4.1. Definitions of each of these independent variables are as follows:

- *Daily Cigarette Intake* self-reported daily cigarette usage by trimester [0(Nonsmoker), 1(1-5 cigarettes), 2(6-10 cigarettes), 3(10-20 cigarettes), 4(21-40 cigarettes), 5(41+ cigarettes)].
- *Gestational Hypertension* or Pregnancy Induced Hypertension is when the mother develops high blood pressure during pregnancy.
- *Hypertension Eclampsia* is the condition of high blood pressure developed during pregnancy as well as proteinuria (an abnormal level of protein in the urine).
- *Risk Factors Determined* includes: Pre-pregnancy Diabetes, Gestational Diabetes, Pre-pregnancy Hypertension, Gestational Hypertension, Hypertension Eclampsia, Infertility Treatment Used, Fertility Enhancing Drugs, or Asst. Reproductive Technology.
- *Payment Method* is the form of payment for delivery/care related to the birth.
- *Term of First Prenatal Visit* is the trimester in which the mother first sought prenatal care during the pregnancy.
- *Number of Prenatal Visits* is the frequency of prenatal visits prior to pregnancy.
- *BMI* is determined by the mother's pre-pregnancy height and weight.
- *Mother's Education Level* is grouped into three categories: high school graduate or below, some college or associates degree, and bachelor's degree or above.
- *Mother's race* includes AIAN (American Indian and Alaskan Native) and NHOPI (Native Hawaiian and Other Pacific Islander).

Table 2 – Descriptive Statistics of Factors Related to Premature Birth

Variable Name	n	Percent	Mean (Min, Max)	St.Dev
First Trimester Daily Cigarettes Intake	1,314,245		0.12(0,5)	0.52
Second Trimester Daily Cigarettes Intake	1,314,245		0.09(0,5)	0.42
Sex (F)	1,314,245	0.488	0.488(0,1)	
Gestational Hypertension (Y)	1,314,245	0.067	0.067(0,1)	
Hypertension Eclampsia (Y)	1,314,245	0.003	0.003(0,1)	
NoRiskFactorsDetermined (Y)	1,314,245	0.849	0.849(0,1)	
PaymentMethod	1,314,245			
Medicaid	509,981	0.388		
Not Reported	9,954	0.008		
Other Payment	53,871	0.041		
Self-Pay	42,658	0.032		
Private Ins	697,781	0.531		
Term of First Prenatal Visit	1,314,245			
First Trimester	1,045,158	0.795		
Second Trimester	214,250	0.163		
Third Trimester	54,837	0.042		
Number of Prenatal Visits	1,314,245	-	11.69(0,98)	3.73
1-6 Visits	94,047	0.072		
7-8 Visits	113,865	0.087		
9-10 Visits	270,172	0.206		
11-16 Visits	753,398	0.573		
17+ Visits	82,763	0.063		
BMI	1,314,245	-	25.77(13,68.9)	6.30
Underweight (<18.5)	66,990	0.051		
Normal (18.5-24.9)	624,706	0.475		
Overweight(25.0-29.9)	351,146	0.267		
Obese Class 1(30.0-34.9)	151,156	0.115		
Obese Class 2+(≥ 35.0)	120,247	0.091		
Mother's Education Level	1,314,245			
High School/GED and Below	466,134	0.355		
Some College or Associates Degree	388,042	0.295		
Bachelor's Degree or Higher	460,069	0.350		
Mother's Age Group	1,314,245			
Teenager (<20 Years Old)	179,836	0.137		
Adult (20-40 Years Old)	1,114,362	0.848		
Adult (40+ Years Old)	20,047	0.015		
Mother's Race Category	1,314,245			
Non-Hispanic White	748,409	0.569		
Hispanic	266,621	0.203		
Non-Hispanic Black	168,501	0.128		
Non-Hispanic Asian	91,995	0.070		
Non-Hispanic Two or More Races	27,719	0.021		
Non-Hispanic AIAN & NHOPI	11,000	0.008		

Table 2 – Descriptive Statistics of Factors Related to Premature Birth
Source: 2014 NCHS Natality Dataset of First-born singleton births (n=1,314,245)

Below in Table 3, each step of the stepwise procedure provides insight into which factors and interactions are most predictive of premature birth. Each step signifies the model that offers the greatest reduction in the AICc score out of the set of all possible models given the inputs. For the final model construction, each of these terms will be added following the stepwise order. Before adding interaction term, each precedent term will be added to the model regardless of stepwise results that allowed for models with missing effects. For example, the interaction term from step three included *Term of First Prenatal Visit,* however this precedent term was not yet entered into the stepwise model. Once all the terms were added from the stepwise output, the model was simplified by iteratively removing the smallest Effects Likelihood Ratio test score until only statistically significant terms remained. Any individual term that was part of a higher ranked interaction term was kept in the model regardless of its statistical significance.

Table 3 - Binary Logistic Stepwise Platform Output for *Premature* Birth

Step	Parameter	L-R Chi Square	R Square	Cumulative Number of parameters	AICc
1	Number Of Prenatal Visits	32547.4	0	5	555392
2	No Risk Factors Determined	8733.518	0.0743	6	514121
3	Term Of First Prenatal Visit*Number Of Prenatal Visits	5366.538	0.084	14	508771
4	Term Of First Prenatal Visit * Mother's Race	1608.128	0.0869	24	507183
5	Mother's Race * Mother's Age	942.5689	0.0886	34	506260
6	Gestational Hypertension	791.0833	0.09	35	505471
7	Hypertension Eclampsia* No Risk Factors Determined	524.8292	0.091	36	504948
8	Number Of Prenatal Visits * Mother's Race	654.1181	0.0921	56	504334
9	BMI Levels * Mother's Age	394.1163	0.0928	64	503956
10	Mother Education Level	278.5666	0.0933	66	503681
11	sex{M-F}	245.9047	0.0938	67	503438
12	Term Of First Prenatal Visit	245.0722	0.0942	69	503196
13	Term Of First Prenatal Visit * Mother Education	266.9608	0.0947	73	502938
14	Payment Method*Number Of Prenatal Visits	244.8525	0.0951	89	502725

Table 3 – Binary Logistic Stepwise Platform Output for *Premature* birth

Finally, for the consideration of providing manageable and easily interpretable odd ratios, multiple interactions involving the same term were removed keeping only the most significant interaction for any single independent variable. Table 4 depicts the full model ranked by descending Likelihood Ratio(L-R) Chi square score. Terms in boxes signify they were removed between the stepwise output Table 3 and the final model in Table 5 due to multiple interactions per factor.

Table 4 – Stepwise model Effects likelihood ratio tests

Source	n	DF	L-R Chi Square	Prob>ChiSq
NoRiskFactorsDetermined	1	1	1692.8847	<.0001
Term of First Prenatal Visit *Num of Prenatal Visits	8	8	1289.19879	<.0001
Num of Prenatal Visits	4	4	1145.53301	<.0001
Gestational Hypertension	1	1	922.815242	<.0001
Hypertension Eclampsia	1	1	515.532293	<.0001
Mother Education	2	2	396.822587	<.0001
Mother Race *Num of Prenatal Visits	20	20	342.25512	<.0001
sex	1	1	278.663625	<.0001
Payment Method*Num of Prenatal Visits	16	16	272.121135	<.0001
Term of First Prenatal Visit*Mother Education	4	4	227.637784	<.0001
BMI Levels*Mother Age	8	8	173.021062	<.0001
Term of First Prenatal Visit	2	2	151.518012	<.0001
Mother Race	5	5	111.684113	<.0001
Payment Method	4	4	70.0972388	<.0001
Mother Age	2	2	49.3030686	<.0001
Term of First Prenatal Visit *Mother Race	10	10	39.276282	<.0001
Mother Race*Mother Age	10	10	26.6377037	0.003
BMI Levels	4	4	10.0537095	0.0395

Table 4 - Stepwise Model Effects likelihood ratio tests scores

The resulting final model is detailed in Chapter 4.1.4 with a full list of parameter estimates, model form, and fit details.

4.1.4 Binary Logistic Regression Parameter Estimates, Model form, and Fit details

Below in table 5 are the model parameters with their respective estimated value and significance test statistics. Nominal variable levels are contained in the bracket and any level (From Table 2) that is not included is the reference cell for that nominal variable.

Table 5 – Binary Logistic Regression Parameter estimate for Premature Birth (<37 weeks)

Model Term - Factor	DF	Estimate	SE	Wald Chi-Square
α Intercept	1	-1.1933	0.0348	1175.7529
β_1 Sex[F]	1	-0.0634	0.0038	278.4912
β_2 Gestational Hypertension [Y]	1	0.2391	0.00792	912.3599
β_3 Hypertension Eclampsia [Y]	1	0.5221	0.0218	575.5787
β_4 No Risk Factors Determined [N]	1	0.2719	0.00634	1840.1606
β_5 PaymentMethod[Medicaid]	1	0.0727	0.0116	38.9936
β_6 PaymentMethod[Not Reported]	1	0.1025	0.0349	8.614
β_7 PaymentMethod[Other Payment]	1	-0.075	0.0179	17.4516
β_8 PaymentMethod[Self Pay]	1	-0.1771	0.0198	79.8452
β_9 Mother Education[Bachelor's and Above]	1	-0.0933	0.00662	198.6692
β_{10} Mother Education[Some College or Associates Degree]	1	0.013	0.00569	5.1942
β_{11} Mother Race [Hispanic]	1	-0.0418	0.0109	14.6268
β_{12} Mother Race [Non-Hispanic AIAN & NHOPI]	1	-0.0623	0.0323	3.7094
β_{13} Mother Race [Non-Hispanic Asian]	1	-0.0823	0.0157	27.4718
β_{14} Mother Race [Non-Hispanic Black]	1	0.2625	0.0113	538.5305
β_{15} Mother Race [Non-Hispanic Two or More Races]	1	-0.0478	0.0228	4.4021
β_{16} Num Of Prenatal Visits[1-6 Visits]	1	0.9623	0.0213	2050.2929
β_{17} Num Of Prenatal Visits[17+ Visits]	1	-0.4856	0.0723	45.1256
β_{18} Num Of Prenatal Visits[7-8 Visits]	1	0.2705	0.0238	128.9899
β_{19} Num Of Prenatal Visits[9-10 Visits]	1	-0.18	0.0266	45.8387
β_{20} Term Of First Prenatal Visit[3rd Trimester Visit]	1	-0.4109	0.0386	113.259
β_{21} Term Of First Prenatal Visit[First Trimester Visit]	1	0.2773	0.0201	190.689
β_{22} BMI[Underweight]	1	0.0605	0.0465	1.6937
β_{23} BMI[Overweight]	1	-0.0481	0.0197	5.9623
β_{24} BMI[Obese Class 1]	1	0.0115	0.0239	0.2305
β_{25} BMI[Obese Class 2+]	1	0.0202	0.0261	0.5992
β_{26} Mother Age [Teenager]	1	-0.1536	0.0166	85.53
β_{27} Mother Age [Over 40]	1	0.334	0.0278	143.9525

Continued Table 5 – Binary Logistic Regression Parameter estimate for Premature Birth (<37 weeks)

Factor	DF	Estimate	SE	Wald Chi-Square
β_{28} Num Of Prenatal Visits*Term Of First Prenatal Visit[1-6 Visits*3rd Trimester]	1	-0.4046	0.0404	100.0624
β_{29} Num Of Prenatal Visits*Term Of First Prenatal Visit[1-6 Visits*First Trimester]	1	0.3577	0.0224	254.881
β_{30} Num Of Prenatal Visits*Term Of First Prenatal Visit[17+ Visits*3rd Trimester]	1	0.3706	0.1415	6.8578
β_{31} Num Of Prenatal Visits*Term Of First Prenatal Visit[17+ Visits*First Trimester]	1	-0.4824	0.0728	43.9148
β_{32} Num Of Prenatal Visits*Term Of First Prenatal Visit[7-8 Visits*3rd Trimester]	1	-0.4127	0.0459	80.8191
β_{33} Num Of Prenatal Visits*Term Of First Prenatal Visit[7-8 Visits*First Trimester]	1	0.4585	0.0246	347.9228
β_{34} Num Of Prenatal Visits*Term Of First Prenatal Visit[9-10 Visits*3rd Trimester]	1	-0.03	0.0518	0.3363
β_{35} Num Of Prenatal Visits*Term Of First Prenatal Visit[9-10 Visits*First Trimester]	1	0.1053	0.027	15.1539
β_{36} BMI*Mother Age[Underweight*Teenager]	1	0.1858	0.0489	14.4351
β_{37} BMI*Mother Age[Underweight*Over 40]	1	-0.1317	0.0912	2.0843
β_{38} BMI*Mother Age[Overweight*Teenager]	1	-0.033	0.0227	2.1192
β_{39} BMI*Mother Age[Overweight*Over 40]	1	0.0394	0.0375	1.1078
β_{40} BMI*Mother Age[Obese Class 1*Teenager]	1	-0.1027	0.0285	12.9868
β_{41} BMI*Mother Age[Obese Class 1*Over 40]	1	0.0872	0.0448	3.7827
β_{42} BMI*Mother Age[Obese Class2+*Teenager]	1	-0.1509	0.0322	21.964
β_{43} BMI*Mother Age[Obese Class 2+*Over 40]	1	0.051	0.048	1.1311

Table 5 - Binary Logistic Regression Parameter Estimates for Premature Birth (<37)

Nominal variables in SAS are coded using effect cell coding [0,1, -1] as depicted in Appendix V for all variables with more than two levels, and are in brackets (Eq 11). Dichotomous terms are modeled using [-1, 1] instead of [0,1] binary/dummy coding, and are in parentheses in the model equation (Eq 11). For example: if the baby's gender is male, the resulting contribution to the model (in linear form) is $\beta_1 * -1(Sex[m]) = 0.063$ while female would be $\beta_1 * 1(Sex[f])$= -0.063.

Model Equation:

$$f(x,\beta,\alpha) = \alpha + \beta_1 \begin{pmatrix} Sex[\mathrm{F}] \\ -(Sex[\mathrm{M}]) \end{pmatrix} + \beta_2 \begin{pmatrix} GestationalHypertension[\mathrm{Y}] \\ -GestationalHypertension[\mathrm{N}] \end{pmatrix}$$

$$+\beta_3 \begin{pmatrix} Hypertension\ Eclampsia[\mathrm{Y}] \\ -\ Hypertension\ Eclampsia[\mathrm{N}] \end{pmatrix} + \beta_4 \begin{pmatrix} \mathrm{NoRiskFactorsDetermined[N]} \\ -\mathrm{NoRiskFactorsDetermined[Y]} \end{pmatrix}$$

$$+ \begin{bmatrix} \beta_5(\mathrm{Payment\ Method[Medicaid]}) + \beta_6(\mathrm{Payment\ Method[Not\ Reported]}) \\ +\beta_7(\mathrm{Payment\ Method[Other\ Payment]}) + \beta_8(\mathrm{Payment\ Method[Self\ Pay]}) \end{bmatrix}$$

$$+ \begin{bmatrix} \beta_9(\mathrm{Mother\ Education\ Groups[Bachelor's\ and\ Above]}) \\ +\beta_{10}(\mathrm{Mother\ Education\ Groups[Some\ College\ or\ Associates]}) \end{bmatrix}$$

$$+ \begin{bmatrix} \beta_{11}(\mathrm{Mother\ Race\ [Hispanic]}) + \beta_{12}(\mathrm{Mother\ Race\ [Non-Hispanic\ AIAN\ \&\ NHOPI]}) \\ +\beta_{13}(\mathrm{Mother\ Race\ [Non-Hispanic\ Asian]}) \\ +\beta_{14}(\mathrm{Mother\ Race\ [Non-Hispanic\ Black]}) + \beta_{15}(\mathrm{Mother\ Race\ [Non-Hispanic\ Two\ or\ More\ Races]}) \end{bmatrix}$$

$$+ \begin{bmatrix} \beta_{16}(\mathrm{Num\ Of\ Prenatal\ Visits[1-6\ Visits]}) + \beta_{17}(\mathrm{Num\ Of\ Prenatal\ Visits\ [7-8\ Visits]}) + \\ \beta_{18}(\mathrm{Num\ Of\ Prenatal\ Visits[9-10\ Visits]}) + \beta_{19}(\mathrm{Num\ Of\ Prenatal\ Visits[17+\ Visits]}) \end{bmatrix}$$

$$+ \begin{bmatrix} \beta_{20}(\mathrm{Term\ Of\ First\ Prenatal\ Visit[First\ Trimester\ Visit]}) \\ +\beta_{21}(\mathrm{Term\ Of\ First\ Prenatal\ Visit[Third\ Trimester\ Visit]}) \end{bmatrix}$$

$$+ \begin{bmatrix} \beta_{22}(\mathrm{BMI[Underweight]}) + \beta_{23}(\mathrm{BMI[Overweight]}) \\ +\beta_{24}(\mathrm{BMI[Obese\ Class\ 1]}) + \beta_{25}(\mathrm{BMI[Obese\ Class\ 2+]}) \end{bmatrix}$$

$$+ \begin{bmatrix} \beta_{26}(\mathrm{Mother\ Age\ [Teenager]}) \\ +\beta_{27}(\mathrm{Mother\ Age\ [40+]}) \end{bmatrix}$$

$$+ \begin{bmatrix} \beta_{28}[\mathrm{1-6\ Visits} * \mathrm{First\ Trimester\ Visit}] + \beta_{29}([\mathrm{1-6\ Visits} * \mathrm{3rd\ Trimester\ Visit}]) + \beta_{30}[\mathrm{7-8\ Visits} * \mathrm{First\ Trimester\ Visit}] + \\ \beta_{31}(\mathrm{7-8\ Visits} * \mathrm{3rd\ Trimester\ Visit}) + \beta_{32}([\mathrm{9-10\ Visits} * \mathrm{First\ Trimester\ Visit}]) + \beta_{33}([\mathrm{9-10\ Visits} * \mathrm{First\ Trimester\ Visit}]) \\ +\beta_{34}([\mathrm{9-10\ Visits} * \mathrm{3rd\ Trimester\ Visit}]) + \beta_{35}([\mathrm{17+\ Visits} * \mathrm{First\ Trimester\ Visit}]) + \beta_{36}([\mathrm{17+\ Visits} * \mathrm{First\ Trimester\ Visit}]) \end{bmatrix}$$

$$+ \begin{bmatrix} \beta_{36}(\mathrm{Underweight} * \mathrm{Teenager}) + \beta_{37}(\mathrm{Underweight} * \mathrm{Over\ 40}) + \beta_{38}(\mathrm{Overweight} * \mathrm{Teenager}) + \\ \beta_{39}(\mathrm{Overweight} * \mathrm{Over\ 40}) + \beta_{40}(\mathrm{Obese\ Class\ 1} * \mathrm{Teenager}) + \beta_{41}(\mathrm{Obese\ Class\ 1} * \mathrm{Over\ 40}) \\ +\beta_{42}(\mathrm{Obese\ Class2} + * \mathrm{Teenager}) + \beta_{43}(\mathrm{Obese\ Class\ 2} + * \mathrm{Over\ 40}) \end{bmatrix}$$

(Eq 11)

Both SAS and JMP model code is included in Appendix II and III with model output reports for each in Appendix IV and V, respectively. Model Fit Detail is summarized below in Table 6.

Table 6 - Fit Detail for Logistic Regression Premature Birth Model

Measure	Training	Validation
n	920,341	393,904
Degrees of Freedom	43	43
Misclassification Rate	0.0894	0.0891
Specificity	91.20%	91.22%
Sensitivity	55.64%	55.39%
Positive Predictive Value	2.42%	2.36%
Negative Predictive Value	99.81%	99.81%
A.U.C.	0.7156	0.7166

Table 6 – Model Fit Detail for Binary Logistic Premature Birth model

Referencing the model fit statistics in Table 6, the first notable finding is that the model preforms similarly between training and validation datasets across all fit statistics. This shows that the model is not over-fitted or over-parametrized. Across each of these datasets, total misclassification rate was 8.94% and 8.91% respectively. Displayed in Appendix IV the misclassifications are depicted visually through sensitivity and specificity curves in the receiver operating characteristic curve (ROC). Finally, looking at the AUC measure, which is the area under the ROC calculated by taking the integral of the combined curves, we see that the model predicts fairly well at 0.72 for both datasets. The AUC measurement can be interpreted as a probability, which may help better understand its usefulness as a single fit statistic. For this model, consider that each observation's predicted probability of *Premature Birth* is sorted from lowest to highest. Then the probability that a randomly selected True Positive has a predicted

probability higher than a randomly selected True Negative is 0.72, an improvement compared to 0.5 that would result from pure random sample without the model's predictive capabilities.

Finally, the below chart depicts the prediction interval generated from the binary logistic regression model crossed by the most significant variable in the model, *Number of Prenatal Visits*.

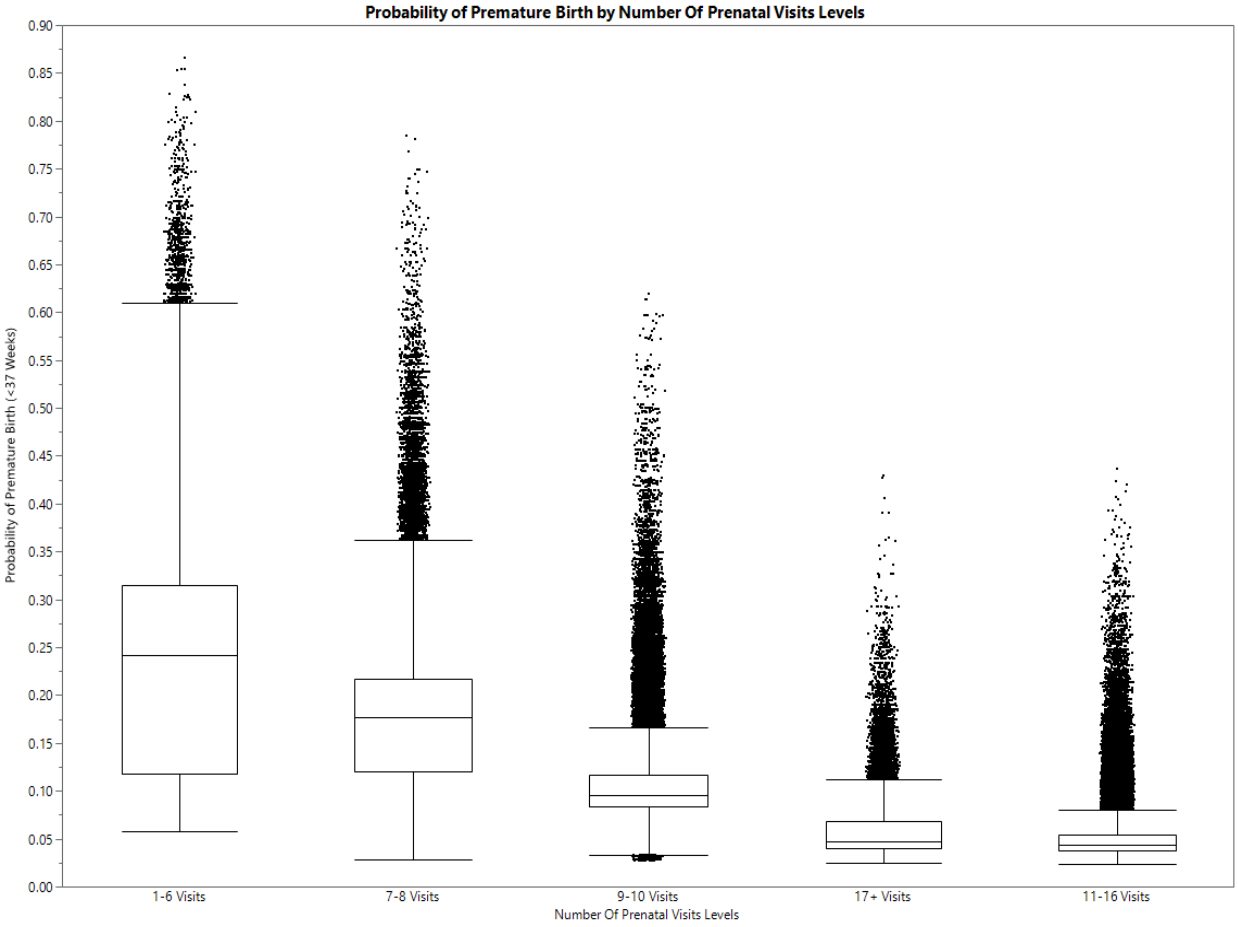


Figure 4 – Probability distribution of Premature Birth by Number of Prenatal Visit levels

4.2 Ordinal Regression Model

The same set of independent variables and second-degree interactions were entered into SAS Institute's JMP Pro 13 Stepwise Regression platform, however the dependent variable has been changed to the ordinal *Prematurity Groups* measure constructed in Chapter 3.2. Recall that level zero represents a *Normal* (≥37 weeks of gestation), level one represent *Premature* (between 32 and 36 weeks), and level two represents *Very Premature* (<32 weeks).

An illustration of the proportional odds model from Chapter 2.3 shows the link functions for the cumulative probabilities used in this thesis model. The Logit link functions for the cumulative probabilities are:

$$P(Normal) = \frac{P(Normal)}{P(Premature)+P(Very\ Prematue)} = (1 - (\frac{e^{(\alpha_1+b_1X)}}{1+e^{(\alpha_1+b_1X)}})) - (\frac{e^{(\alpha_2+b_1X)}}{1+e^{(\alpha_2+b_1X)}})$$

$$P(Premature) = \frac{P(Normal)+P(Premature)}{P(Very\ Premature)} = (\frac{e^{(\alpha_1+b_1X)}}{1+e^{(\alpha_1+b_1X)}}) - (\frac{e^{(\alpha_2+b_1X)}}{1+e^{(\alpha_2+b_1X)}})$$

$$P(Very\ Premature) = P(Very\ Premature) = (\frac{e^{(\alpha_2+b_1X)}}{1+e^{(\alpha_2+b_1X)}}).$$

Note the subscripts on α indicate two intercepts in the log-odds scale of the model representing the level shift from the base response level, in this case *Normal* birth, while b_1X parameters are shared between each logit. The model equation (Eq 12) takes the place b_1X, in all these cumulative probability equations.

It is important to note that modeling *Very Premature* will pose a challenge to fit a model compared to *Premature* from the binary logistic model due to its rare occurrence in the population and the mechanics of logistic regression. *Very Premature* represents only 20,868 of the original 1,329,778 observations or 1.57% of the sample. Univariate data preparations showed small cell size in the most highly correlated variable, *Number of Prenatal Visits,* for the *Very Premature*

category as small as 132 observations. In addition, all interactions resulting in a cell size of ten or fewer observations were excluded. Because of this, the construction of this model is used only as an example of the proportional odds model form.

Table 7 – Ordinal Logistic Regression Model Parameter Estimates

Term	Factor	Estimate	Std Error	WALD Chi-Square
α_2	Intercept [2]	-4.1628	0.0189	48327
α_1	Intercept [1]	-2.2103	0.0172	16449
β_1	Sex[F]	-0.0634	0.0038	282.09
β_2	Gestational Hypertension [Y]	0.4577	0.0156	858.84
β_3	Hypertension Eclampsia [Y]	1.0555	0.0417	639.87
β_4	No Risk Factors Determined [N]	0.2818	0.0063	2025
β_5	PaymentMethod[Medicaid]	0.0737	0.0116	40.62
β_6	PaymentMethod[Not Reported]	0.1032	0.0347	8.86
β_7	PaymentMethod[Other Payment]	-0.073	0.0178	16.76
β_8	PaymentMethod[Self Pay]	-0.1845	0.0197	87.75
β_9	Mother Education[Bachelor's and Above]	-0.1007	0.0066	235.46
β_{10}	Mother Education[Some College or Associates Degree]	0.0176	0.0056	9.72
β_{11}	Mother Race [Hispanic]	-0.0389	0.0109	12.8
β_{13}	Mother Race [Non-Hispanic AIAN & NHOPI]	-0.0896	0.0323	7.71
β_{13}	Mother Race [Non-Hispanic Asian]	-0.0883	0.0156	32
β_{14}	Mother Race [Non-Hispanic Black]	0.2898	0.0112	668.51
β_{15}	Mother Race [Non-Hispanic Two or More Races]	-0.0475	0.0227	4.38
β_{16}	Num Of Prenatal Visits[1-6 Visits]	1.3595	0.0094	20829
β_{17}	Num Of Prenatal Visits[7-8 Visits]	0.611	0.0088	4841.4
β_{18}	Num Of Prenatal Visits[9-10 Visits]	-0.0877	0.0076	134.65
β_{19}	Num Of Prenatal Visits[17+ Visits]	-0.9292	0.0147	3998
β_{20}	Term Of First Prenatal Visit[First Trimester Visit]	0.5706	0.0082	4844.4
β_{21}	Term Of First Prenatal Visit[3rd Trimester Visit]	-0.8382	0.0131	4072.3
β_{22}	BMI[Underweight]	0.0454	0.0135	11.31
β_{23}	BMI[Overweight]	-0.0708	0.0075	89.92
β_{24}	BMI[Obese Class 1]	0.011	0.0096	1.33
β_{25}	BMI[Obese Class 2+]	0.1048	0.0101	107.33
β_{26}	Mother Age [Teenager]	-0.1324	0.0114	133.98
β_{27}	Mother Age [Over 40]	0.3395	0.0176	371.3

Table 7 - Ordinal Logistic Regression Model Parameter Estimates

In JMP, two linear predictors are calculated using α_1 *and* α_2 for the response level *Premature* and *Very Premature*, respectively. To get the third category probability the two linear forms are converted through the three logit link functions illustrated at the beginning of section 4.2. The final model equation (Eq 12):

$$
\begin{aligned}
f(x,\beta) = \; & \beta_1 \begin{pmatrix} Sex[\mathrm{F}] \\ -(Sex[\mathrm{M}]) \end{pmatrix} + \beta_2 \begin{pmatrix} GestationalHypertension[\mathrm{Y}] \\ -GestationalHypertension[\mathrm{N}] \end{pmatrix} \\
& + \beta_3 \begin{pmatrix} Hypertension\ Eclampsia[\mathrm{Y}] \\ -\ Hypertension\ Eclampsia[\mathrm{N}] \end{pmatrix} + \beta_4 \begin{pmatrix} \mathrm{NoRiskFactorsDetermined[N]} \\ -\mathrm{NoRiskFactorsDetermined[Y]} \end{pmatrix} \\
& + \begin{bmatrix} \beta_5(\mathrm{Payment\ Method[Medicaid]}) + \beta_6(\mathrm{Payment\ Method[Not\ Reported]}) \\ +\beta_7(\mathrm{Payment\ Method[Other\ Payment]}) + \beta_8(\mathrm{Payment\ Method[Self\ Pay]}) \end{bmatrix} \\
& + \begin{bmatrix} \beta_9(\mathrm{Mother\ Education\ Groups[Bachelor's\ and\ Above]}) \\ +\beta_{10}(\mathrm{Mother\ Education\ Groups[Some\ College\ or\ Associates]}) \end{bmatrix} \\
& + \begin{bmatrix} \beta_{11}(\mathrm{Mother\ Race\ [Hispanic]}) + \beta_{12}(\mathrm{Mother\ Race\ [Non-Hispanic\ AIAN\ \&\ NHOPI]}) \\ +\beta_{13}(\mathrm{Mother\ Race\ [Non-Hispanic\ Asian]}) \\ +\beta_{14}(\mathrm{Mother\ Race\ [Non-Hispanic\ Black]}) + \beta_{15}(\mathrm{Mother\ Race\ [Non-Hispanic\ Two\ or\ More\ Races]}) \end{bmatrix} \\
& + \begin{bmatrix} \beta_{16}(\mathrm{Num\ Of\ Prenatal\ Visits[1-6\ Visits]}) + \beta_{17}(\mathrm{Num\ Of\ Prenatal\ Visits\ [7-8\ Visits]}) + \\ \beta_{18}(\mathrm{Num\ Of\ Prenatal\ Visits[9-10\ Visits]}) + \beta_{19}(\mathrm{Num\ Of\ Prenatal\ Visits[17+\ Visits]}) \end{bmatrix} \\
& + \begin{bmatrix} \beta_{20}(\mathrm{Term\ Of\ First\ Prenatal\ Visit[First\ Trimester\ Visit]}) \\ +\beta_{21}(\mathrm{Term\ Of\ First\ Prenatal\ Visit[Third\ Trimester\ Visit]}) \end{bmatrix} \\
& + \begin{bmatrix} \beta_{22}(\mathrm{BMI[Underweight]}) + \beta_{23}(\mathrm{BMI[Overweight]}) \\ +\beta_{24}(\mathrm{BMI[Obese\ Class\ 1]}) + \beta_{25}(\mathrm{BMI[Obese\ Class\ 2+]}) \end{bmatrix} \\
& + \begin{bmatrix} \beta_{26}(\mathrm{Mother\ Age\ [Teenager]]}) \\ +\beta_{27}(\mathrm{Mother\ Age\ [40+]}) \end{bmatrix}
\end{aligned}
$$

(Eq 12)

Finally, the below chart depicts the prediction interval generated from the ordinal regression model crossed by the most significant term in the model. Blue boxes depict the spread of predicted values of *Premature birth* while the red boxes show *Very Premature birth*.

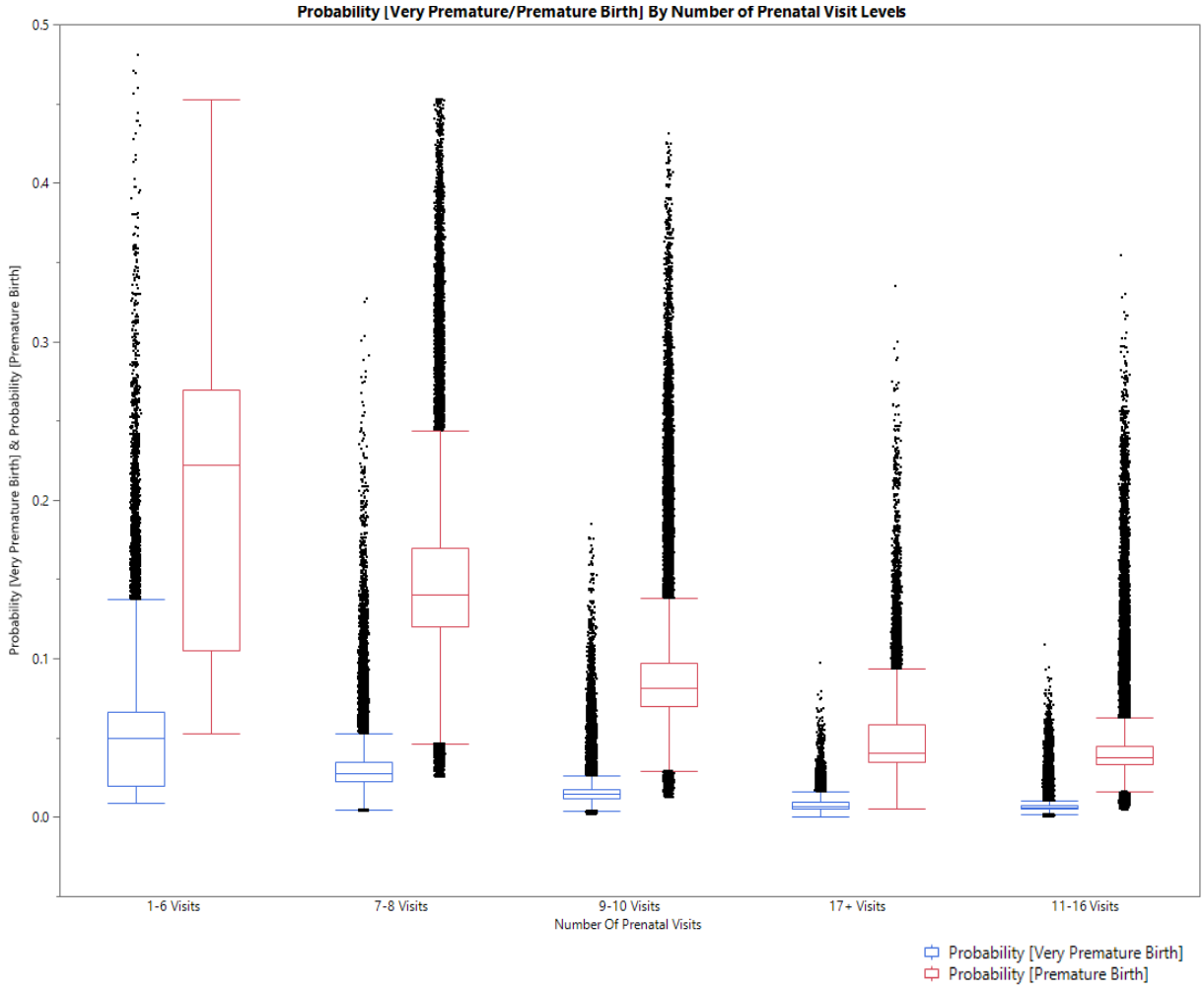


Figure 5 – Ordinal Prediction Intervals by Number of Prenatal visits groups

Chapter 5 - Reported Findings

5.1 Univariate confirmation of Risk Factors in 2014 Natality Dataset

This section will summarize overall findings from univariate data preparation of the full dataset. Recall the World Health Organization's (WHO) list of risk factors presented in Chapter 1.2. From the full dataset of 3,845,514 births in 2014 with a reported gestational period, the following prevalence rates of premature birth for the risk factors listed were verified. Nearly 57% of all twins and 94% of all triplets are born premature (Table 1). Mothers that had a prior premature birth showed increased prevalence of prematurity at 29.4%. Mothers that did not have a prior premature birth had prevalence of prematurity at just 10.8%. During the observation reduction process in Chapter 3.1.1, an additional influential group of observations was identified. Within the first time singleton mother sample, a mother that reported no prenatal care prior to delivery showed 22.9% prevalence of premature birth. Only first time singleton birth mothers with at least one prenatal visit prior to delivery were included in the final dataset for the multivariate model; all observations from plural births, prior preterm births, and mothers that did not attend prenatal care prior to delivery as discussed in this section are excluded from findings reported in the next section.

5.2 Multivariate Model Findings (Binary Logistic Model) and Discussion

Multivariate analysis provides evidence that frequent and routine prenatal care has a strong relationship with lower incidence of *Premature* birth. For a routine pregnancy, standard medical practice schedules mothers for monthly checkups until week 28, at which point checkups are scheduled bi-weekly. At week 36, weekly appointments are scheduled until delivery. Under this standard practice, a mother who begins care in the first trimester will attend a minimum of 12 prenatal visits for a full-term (40 week) pregnancy. If she begins care in the second or third

trimester, standard practice schedules her to attend 10 or 8 prenatal visits, respectively. Mothers with risks or complications will often require more frequent visits than this baseline standard; in the most extreme circumstances the mother may be admitted to the hospital at the discretion of her obstetrician. In the modeling dataset there were 44,043 (3.3%) mothers that attended nineteen or more prenatal visits with a prevalence of prematurity of just 7.20%, which is below the rate for all mothers, at 10.34% premature.

Consider the odds ratios (recalling the definition from Chapter 2.2) in *Table 8*, a pattern emerges for the interaction between *Term of First Prenatal Visit* and *Number of Prenatal Visits*. Mothers with 1-6 visits compared to the baseline 11-16 visits show increasing risk of premature birth when the first visit is earlier in the pregnancy. The baseline group of 11-16 visits was chosen because it contains the mean number of visits (11.24 visits) and accounts for 57.3% of the sample. Also note that the 11-16 visit category has the lowest prematurity rate (5.22%) compared to all other *Number of Prenatal Visits* levels.

The model result illustrates that attending fewer than the overall standard number of visits discussed above dramatically increases the risk of prematurity. To explore this idea further, consider different baselines on the Odds Ratios for each individual *Term of First Prenatal Visit* level, that is based on the minimum standard at that level discussed earlier in this section. The following list gives the odds ratio (with 95% CI) for mothers that attended 1-6 visits compared to the minimal standard prenatal care regiment based on the trimester of the first visit.

- Prenatal care starting in 1st Trimester 1-6 Visits vs. 11-16 Visits 10.24 (9.93, 10.57)
- Prenatal care starting in 2nd Trimester 1-6 Visits vs. 9-10 Visits 3.54 (3.38, 3.71)
- Prenatal care starting in 3rd Trimester 1-6 Visits vs. 7-8 Visits 2.01 (1.82, 2.23)

Compared to a fixed baseline group *Number of Prenatal Visits* of 11-16 visits, the odds of premature birth increases from 1.91 (1.61, 2.26) in the third trimester to 10.24 (9.93, 10.57) times greater if care began in the first trimester. This same pattern also extends to the other *Number of Prenatal Visits* levels as well, as illustrated in Figure 6 below:

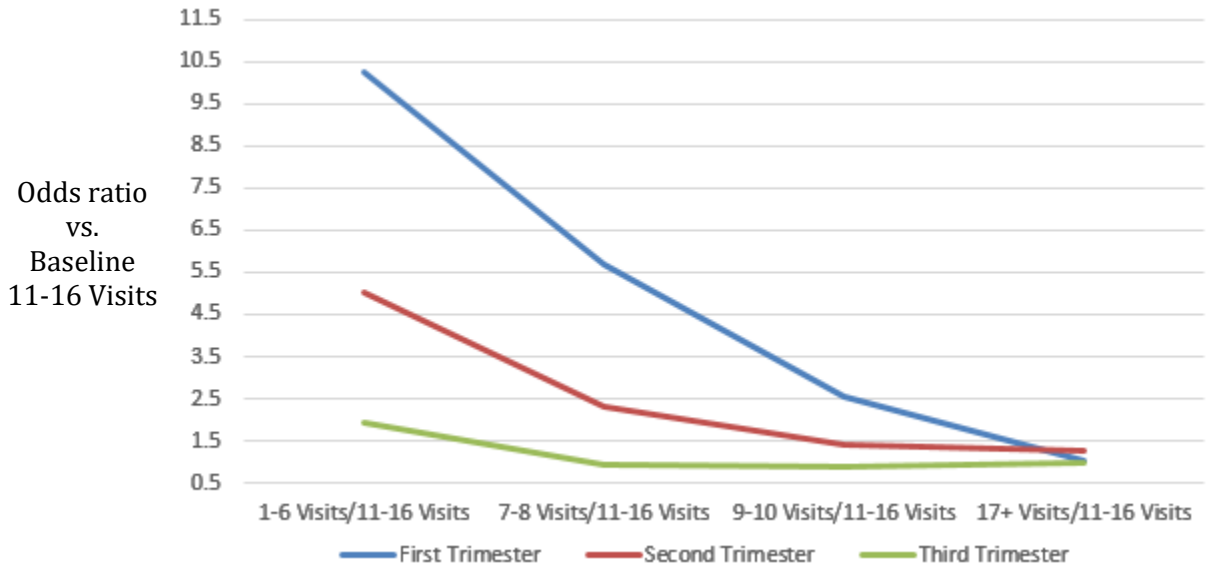


Figure 6 – Odds Ratios for each level of Number of Prenatal Visits Levels by Term of First Prenatal Visit

It appears that the odds of prematurity are greater for each *Number of Prenatal Visits* level for mothers that began prenatal care earlier in their pregnancy, compared to the baseline *Number of Prenatal Visits* [11-16].

The multivariate model had an interesting effect on the *Payment Method* variables compared to preliminary univariate findings. When modeled alone, all levels of *Payment Method* showed increased prevalence of prematurity compared to Private Insurance (Odds Ratio and 95% CI for Medicaid/Private Insurance 1.32 (1.30, 1.34)). However, when demographic and prenatal care factors are included in the model, as in the final multivariate model, Medicaid becomes

statistically insignificant compared to private insurance (Table 8 – Odds ratio confidence interval for Medicaid contains 1). Another finding that may have been missed without an interaction term involves *BMI Level* and *Mother Age*. Both obese teenagers and underweight mothers over forty years old are conversely at lower risk of prematurity compared to their normal weight peers (Table 8). This model result shows that the Mayo Clinic's risk factor from Chapter 1.2 applies only to mothers aged 20 to 40 years old, albeit the majority of mothers are in that age category.

Other notable odds ratios include: if a mother was identified with *Hypertension Eclampsia* her odds are 2.84 (2.61, 3.09) times greater to deliver prematurely vs mothers without *Hypertension Eclampsia*. If *No Risk Factor Determined* is "No", meaning that a risk factor was identified, the odds of prematurity increases by 1.72 (1.68, 1.77) compared to *No Risk Factor Determined* is "Yes". The full set of odds ratios from the final model can be found in the following Table 8.

Table 8 – Multivariate Odds Ratios for Risk of Premature Birth (<37)

Factor	O.R. (95% CI)
Gestational Hypertension [Y/N]*	1.61 (1.56, 1.66)
Hypertension Eclampsia [Y/N]*	2.84 (2.61, 3.09)
No Risk Factor Determined [N/Y]*	1.72 (1.68, 1.77)
SEX [Male/Female]*	1.14 (1.12, 1.15)
Payment Method	
Medicaid/Private Insurance	1 (0.98, 1.02)
Not Reported/Private Insurance	1.03 (0.94, 1.12)
Private Insurance	
Other Payment/Private Insurance*	0.86 (0.83, 0.89)
Self-Pay/Private Insurance*	0.78 (0.74, 0.81)
BMI – Teenagers	
Underweight/Normal Range*	1.21 (1.13, 1.29)
Normal Range	
Overweight/Normal Range*	0.87 (0.83, 0.91)
Obese/Normal Range*	0.86 (0.81, 0.92)
Obese Class 1+/Normal Range*	0.83 (0.77, 0.9)
BMI - Adults (20-40)	
Underweight/Normal Range*	1.11 (1.07, 1.16)
Normal Range	
Overweight/Normal Range*	1.05 (1.02, 1.07)
Obese/Normal Range*	1.13 (1.11, 1.16)
Obese Class 1+/Normal Range*	1.25 (1.21, 1.28)
BMI - Adults (40+)	
Underweight/Normal Range	1.02 (0.73, 1.43)
Normal Range	
Overweight/Normal Range	1.09 (0.96, 1.23)
Obese/Normal Range*	1.21 (1.04, 1.41)
Obese Class 1+/Normal Range	1.18 (1, 1.38)
Mother's Age (Underweight BMI)	
Teenager/Adult (20-40)*	1.31 (1.22, 1.4)
Over 40/Adult (20-40)*	1.55 (1.11, 2.15)
Mother's Age (Normal BMI)	
Teenager/Adult (20-40)*	1.2 (1.16, 1.24)
Over 40/Adult (20-40)*	1.69 (1.56, 1.83)

Table 8 - Continued - Multivariate Odds Ratios for Risk of Premature Birth (<37)

Factor	O.R. (95% CI)
Mother's Age (Overweight BMI)	
Teenager/Adult (20-40)	1 (0.96, 1.04)
Over 40/Adult (20-40)*	1.75 (1.59, 1.93)
Mother's Age (Obese Class 1 BMI)	
Teenager/Adult (20-40)*	0.91 (0.86, 0.97)
Over 40/Adult (20-40)*	1.8 (1.57, 2.05)
Mother's Age (Obese Class 2+ BMI)	
Teenager/Adult (20-40)*	0.8 (0.74, 0.87)
Over 40/Adult (20-40)*	1.59 (1.38, 1.84)
Mother's Race/Ethnicity	
Hispanic/Non-Hispanic White	0.99 (0.97, 1.01)
Non-Hispanic Two or More Races/Non-Hispanic White	0.98 (0.93, 1.03)
Non-Hispanic Asian/Non-Hispanic White*	0.95 (0.92, 0.98)
Non-Hispanic White	
Non-Hispanic AIAN & NHOPI/Non-Hispanic White	0.97 (0.9, 1.04)
Non-Hispanic Black/Non-Hispanic White*	1.34 (1.31, 1.37)
Number of Prenatal Visits (First Trimester Start)	
1-6 Visits/11-16 Visits*	10.24 (9.93, 10.57)
7-8 Visits/11-16 Visits*	5.67 (5.52, 5.82)
9-10 Visits/11-16 Visits*	2.54 (2.49, 2.59)
11-16 Visits	
17+ Visits/11-16 Visits*	1.04 (1, 1.08)
Number of Prenatal Visits (Second Trimester Start)	
1-6 Visits/11-16 Visits*	5.02 (4.8, 5.26)
7-8 Visits/11-16 Visits*	2.29 (2.19, 2.41)
9-10 Visits/11-16 Visits*	1.42 (1.35, 1.49)
11-16 Visits	
17+ Visits/11-16 Visits*	1.26 (1.12, 1.42)
Number of Prenatal Visits (Third Trimester Start)	
1-6 Visits/11-16 Visits*	1.91 (1.62, 2.26)
7-8 Visits/11-16 Visits	0.95 (0.79, 1.15)
9-10 Visits/11-16 Visits	0.89 (0.72, 1.09)
11-16 Visits	
17+ Visits/11-16 Visits	0.98 (0.57, 1.67)
Mother's Education level	
High School Graduate or GED	
Some College or Associates Degree / High School Grad or GED*	0.94 (0.92, 0.95)
Bachelor's and Above / High School Graduate or GED*	0.9 (0.88, 0.92)

Table 8 – Multivariate Odds Ratios for Risk of Premature Birth
Modeled from 2014 NCHS Natality Dataset of first-born singleton births n=1,314,245.
*Premature Birth derived from (*GESTREC10*) with a birth week prior to the 37th week of gestation.*
**Odds Ratios is statistically significant (>0.05).*

Chapter 6 – Summary Thesis Remarks and Future Study

This thesis provided the opportunity to become familiar with clinical data and reporting in an active field of medical research. Data decisions are crucial and their impact can be substantial. Clinical research that strives for unbiased interpretation of factors, must be based on the intimate knowledge of the data and collection process. Throughout the process of data exploration, cleansing, model building, model evaluation, and finally the reporting of findings, the importance of data knowledge was affirmed. In many cases there is no perfect model and rational decisions must be made by researchers with noble interests of reporting unbiased findings. My personal takeaway from the analysis is quite simple. To reduce premature birth, make the importance of prenatal care known, readily available, and accessible to mothers.

For future study, techniques such as oversampling could be explored to achieve more sensitive models for *Very Premature* (1.62% of sample) as predicting rare events with logistic regression often produces low probability by inherent design. In addition, other Ordinal Regression and Multinomial Regression model forms could be explored that allow for unequal slopes between response levels. Finally, if more granular data were available, the further exploration of the relationship between the frequency of prenatal care and its timing during the stages of pregnancy could be researched.

6.1 Limitations

It is important to note that there are limitations to observational studies compared to randomized controlled trials (RCT), which are considered the "gold standard". Since observational studies are uncontrolled by nature the finding can be biased and subject to confounding. While observational studies are efficient and can identify relationships, they cannot prove causality. In order to determine if the *Number of Prenatal Visits* is the cause of

Premature birth, a designed experiment with test and control groups would be required. However, these experiments are time consuming, expensive, and possibly unethical as it may put the health of babies and mothers at risk. It is also important to note that there may be other factors not captured in the dataset that may explain *Premature* birth more accurately, for example stressful life events or physical trauma.

References

1. Liu, L., Oza, S., Hogan, D., Perin, J., Rudan, I., Lawn, J., Cousens, S., Mathers, C. and Black, R. (2014). *Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis*. [online] The Lancet. Available at: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)61698-6/fulltext [Accessed 17 Jul. 2019].
2. Kochanek, K., Murphy, S., Xu, J. and Arias, E. (2019). *Products - Data Briefs - Number 293 - December 2017*. [online] Cdc.gov. Available at: https://www.cdc.gov/nchs/products/databriefs/db293.htm [Accessed 17 Jul. 2019].
3. Marchofdimes.org. (2013). *Long-term health effects of premature birth*. [online] Available at: https://www.marchofdimes.org/complications/long-term-health-effects-of-premature-birth.aspx [Accessed 17 Jul. 2019].
4. Mayo Clinic. (2018). *Premature birth - Symptoms and causes*. [online] Available at: https://www.mayoclinic.org/diseases-conditions/premature-birth/symptoms-causes/syc-20376730 [Accessed 17 Jul. 2019].
5. O'Connor, J. and Robertson, E. (1999). *Adrien-Marie Legendre (1752-1833)*. [online] Www-history.mcs.st-and.ac.uk. Available at: http://www-history.mcs.st-and.ac.uk/Biographies/Legendre.html [Accessed 17 Jul. 2019].
6. Alan Agresti. *Categorical Data Analysis - Second Edition*. Hoboken, New Jersey: John Wiley & Sons, Incorporated;2002.
7. Long-term health effects of premature birth. March of Dimes Organization. https://www.marchofdimes.org/complications/long-term-health-effects-of-premature-birth.aspx. Publication date unavailable. Updated October 2013. Accessed July 18, 2019.
8. Cdc.gov. (2015). *Data Access - Public-Use Data Files and Documentation*. [online] Available at: https://www.cdc.gov/nchs/data_access/ftp_data.htm [Accessed 17 Jul. 2019].
9. Wonder.cdc.gov. (2017). *Fetal Death Records - Online Database Help*. [online] Available at: https://wonder.cdc.gov/wonder/help/fetal-deaths.html [Accessed 17 Jul. 2019].
10. Hamilton, Ph.D., B., Martin, M.P.H, J., Osterman, M.H.S., M., Curtin, M.A., S. and Mathews, M.S., T. (2015). *National Vital Statistics Reports*. [online] Cdc.gov. Available at: https://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_12.pdf [Accessed 17 Jul. 2019].
11. Jmp.com. (2019). *Stepwise Regression Control Panel*. [online] Available at: https://www.jmp.com/support/help/14-2/stepwise-regression-control-panel.shtml [Accessed 17 Jul. 2019].

Additional References:
Peter McCullagh. *Regression Models for Ordinal Data*. Journal of the Royal Statistical Society. Vol. 42, No. 2; 1980.
Ann. R Cannon, George W. Cobb, et.al. *STATS2: Building Models for a World of Data*, W.H. Freeman and Company, New York; 2013.

Dataset:
Natality File. Centers for Disease Control and Prevention.
ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/DVS/natality/Nat2014us.zip. 2014

Appendix I – Data Preparation – Key Variables

Original Database n	3,998,175
Less Missing Dependent variable (*GESTREC10*)	3,303
Less plural births (*dplural* >1)	140,115
Total Remaining Sample	**3,854,757**

Characteristic (variable name)	% Very Premature (<32 weeks)	% Premature (<37 weeks)	% of Sample	N
Paternity Acknowledgement (mar_p)				
Yes	1.83%	10.93%	27.41%	1,056,625
No	2.81%	14.15%	11.58%	446,402
Unknown	3.34%	15.21%	0.15%	5,831
Not Applicable	1.11%	8.11%	57.31%	2,209,078
Total Reported	1.52%	9.65%	96.45%	3,717,936
Missing Records (coded .)	1.47%	8.82%	3.55%	136,821
Total	1.52%	9.62%	100%	3,854,757
Mother's Age (MAGER9)				
1 Under 15 years	5.19%	19.99%	0.07%	2,737
2 15-19 years	2.21%	12.05%	6.36%	245,282
3 20-24 years	1.63%	10.06%	22.38%	862,733
4 25-29 years	1.37%	8.85%	28.84%	1,111,562
5 30-34 years	1.31%	8.74%	26.93%	1,037,961
6 35-39 years	1.63%	10.35%	12.56%	483,971
7 40-44 years	2.15%	13.08%	2.69%	103,738
8 45-49 years	2.67%	16.33%	0.16%	6,300
9 50-54 years	3.59%	19.45%	0.01%	473
Total	1.52%	9.62%	100%	3,854,757
Father's Age (FAGEREC11)				
1 Under 15 years	6.32%	17.47%	0.01%	269
2 15-19 years	2.11%	11.97%	2.17%	83,542
3 20-24 years	1.57%	9.96%	12.65%	487,622
4 25-29 years	1.29%	8.58%	21.90%	844,114
5 30-34 years	1.14%	8.13%	25.60%	986,811
6 35-39 years	1.25%	8.74%	15.65%	603,287
7 40-44 years	1.51%	9.99%	6.46%	248,867
8 45-49 years	1.80%	11.15%	2.14%	82,542
9 50-54 years	2.01%	11.90%	0.69%	26,723
10 55-98 years	2.13%	13.09%	0.30%	11,620
11 Not Stated/Missing	1.34%	8.97%	12.44%	479,360
Total	1.52%	9.62%	100.00%	3,854,757
Month Prenatal Care Began (PRECARE5)				
1 1st to 3rd month	1.24%	8.58%	71.00%	2,737,008
2 4th to 6th month	1.98%	12.07%	16.18%	623,734
3 7th to final month	1.31%	10.60%	4.20%	161,787
4 No prenatal care	6.72%	23.69%	1.45%	55,704
5 Unknown or not stated	1.52%	13.14%	3.62%	139,703
Total Reported	1.52%	9.65%	96.45%	3,717,936
Missing Records (coded .)	1.47%	8.82%	3.55%	136,821
Total	1.52%	9.62%	100%	3,854,757

Number of Prenatal Visits Recode (PREVIS_REC)	% Very Premature (<32 weeks)	% Premature (<37 weeks)	% of Sample	N
0 No visits	6.77%	23.79%	1.47%	56,753
01 1 to 2 visits	7.41%	23.23%	1.12%	43,195
02 3 to 4 visits	8.27%	22.59%	2.44%	93,935
03 5 to 6 visits	6.34%	20.75%	4.98%	191,989
04 7 to 8 visits	2.62%	16.35%	9.16%	352,943
05 9 to 10 visits	1.05%	10.57%	20.78%	800,843
06 11 to 12 visits	0.50%	6.28%	25.23%	972,563
07 13 to 14 visits	0.37%	4.74%	17.29%	666,577
08 15 to 16 visits	0.47%	5.68%	9.43%	363,602
09 17 to 18 visits	0.45%	5.94%	2.42%	93,224
10 19 or more visits	0.78%	8.66%	3.00%	115,496
11 Unknown/Not Stated	1.52%	14.45%	2.69%	103,637
Total	1.52%	9.62%	100%	3,854,757
Cigarettes 1st Trimester Recode (cig1_r)				
0 Nonsmoker	1.44%	9.31%	86.81%	3,346,425
1 1-5	2.38%	12.86%	2.56%	98,813
2 6-10	2.31%	13.01%	3.09%	119,072
3 11-20	2.28%	13.95%	1.89%	72,947
4 21-40	2.73%	15.28%	0.17%	6,625
5 41 or more	2.39%	15.45%	0.03%	1,256
6 Unknown/Not Stated	2.00%	10.53%	1.89%	72,798
Total Reported	1.52%	9.65%	96.45%	3,717,936
Missing Records (coded .)	1.47%	8.82%	3.55%	136,821
Total	1.52%	9.62%	100%	3,854,757
Cigarettes 2nd Trimester Recode (cig2_r)				
0 Nonsmoker	1.45%	9.33%	87.92%	3,389,273
1 1-5	2.44%	13.34%	2.53%	97,531
2 6-10	2.21%	13.31%	2.84%	109,576
3 11-20	2.39%	14.70%	1.15%	44,199
4 21-40	3.38%	17.30%	0.08%	3,139
5 41 or more	2.45%	15.42%	0.02%	856
6 Unknown or not stated	2.02%	10.55%	1.90%	73,362
Total Reported	1.52%	9.65%	96.45%	3,717,936
Missing Records (coded .)	1.47%	8.82%	3.55%	136,821
Total	1.52%	9.62%	100%	3,854,757
Gestational Hypertension (RF_GHYPE)				
No	1.41%	9.05%	91.54%	3,528,643
Yes	3.67%	21.03%	4.75%	183,004
Unknown	3.61%	13.60%	0.16%	6,289
Total Reported	1.52%	9.65%	96.45%	3,717,936
Missing Records (coded .)	1.47%	8.82%	3.55%	136,821
Total	1.52%	9.62%	100.00%	3,854,757
Hypertension Eclampsia (RF_EHYPE)				
No	1.50%	9.58%	96.06%	3,702,789
Yes	9.53%	36.70%	0.23%	8,858
Unknown	3.61%	13.60%	0.16%	6,289
Total Reported	1.52%	9.65%	96.45%	3,717,936
Missing Records (coded .)	1.47%	8.82%	3.55%	136,821

Total	1.52%	9.62%	100%	3,854,757

Prior Pre-Term delivery (RF_PPTERM)	% Very Premature (<32 weeks)	% Premature (<37 weeks)	% of Sample	N
Yes	5.07%	26.81%	2.62%	101,139
No	1.42%	9.16%	93.66%	3,610,508
Unknown	3.61%	13.60%	0.16%	6,289
Total Reported	1.52%	9.65%	96.45%	3,717,936
Missing Records (coded .)	1.47%	8.82%	3.55%	136,821
Total	1.52%	9.62%	100%	3,854,757
No Risk Factors (NO_RISKS)				
1 - No Risks	1.25%	8.11%	69.39%	2,674,795
0 - Risk Factor Reported	2.21%	13.59%	26.90%	1,036,852
9 - Not Reported	3.61%	13.60%	0.16%	6,289
Total Reported	1.52%	9.65%	96.45%	3,717,936
Missing Records (coded .)	1.47%	8.82%	3.55%	136,821
Total	1.52%	9.62%	100%	3,854,757
Payment Type Recode				
1 Medicaid	1.90%	11.41%	41.94%	1,616,641
2 Private Insurance	1.15%	7.97%	45.32%	1,746,911
3 Self Pay	1.70%	9.45%	4.11%	158,341
4 Other	1.51%	10.04%	4.21%	162,447
9 Unknown	2.10%	11.10%	0.87%	33,596
Total Reported	1.52%	9.65%	96.45%	3,717,936
Missing Records (coded .)	1.47%	8.82%	3.55%	136,821
Total	1.52%	9.62%	100%	3,854,757
Prior Dead - Recode (PRIORDEAD)				
No Prior Dead	1.49%	9.52%	98.36%	3,791,610
Yes Prior Dead	4.00%	17.40%	1.18%	45,382
No Reported	2.61%	11.46%	0.46%	17,765
Total	1.52%	9.62%	100%	3,854,757
Gender of Baby (SEX)				
Male	1.59%	10.11%	51.20%	1,973,610
Female	1.45%	9.11%	48.80%	1,881,147
Total	1.52%	9.62%	100%	3,854,757
Mother's Education Level (Meduc)				
1 8th grade or less	1.73%	11.55%	3.51%	135,307
2 9th through 12th grade	2.14%	12.78%	11.05%	425,822
3 High school graduate or GED	1.84%	10.94%	24.10%	929,075
4 Some college credit	1.60%	9.84%	20.47%	789,070
5 Associate degree	1.35%	9.09%	7.71%	297,041
6 Bachelor’s degree	0.95%	7.11%	18.19%	701,121
7 Master’s degree	0.92%	6.95%	8.05%	310,160
8 Doctorate	0.84%	6.97%	2.30%	88,653
9 Unknown	1.80%	9.43%	4.63%	178,508
Total	1.52%	9.62%	100%	3,854,757

Mother's BMI (BMI)	% Very Premature (<32 weeks)	% Premature (<37 weeks)	% of Sample	N
1 Underweight <18.5	1.78%	11.58%	3.54%	136,308
2 Normal 18.5-24.9	1.28%	8.93%	42.70%	1,646,058
3 Overweight 25.0-29.9	1.44%	9.35%	23.76%	915,926
4 Obesity I 35.0-34.9	1.70%	10.19%	12.63%	487,002
5 Obesity II 35.0-39.9	1.91%	10.82%	6.04%	232,970
6 Extreme Obesity III ≥ 40.0	2.19%	11.69%	4.18%	161,039
9 Unknown or not stated	2.74%	12.10%	3.60%	138,633
Total Reported	1.52%	9.65%	96.45%	3,717,936
Missing Records (coded .)	1.47%	8.82%	3.55%	136,821
Total	1.52%	9.62%	100%	3,854,757
Mother's Race (mracehisp)				
1 Non-Hispanic White (only)	1.16%	8.30%	51.09%	1,971,342
2 Non-Hispanic Black (only)	3.14%	14.59%	13.60%	524,676
3 Non-Hispanic AIAN (only)	1.82%	12.57%	0.82%	31,715
4 Non-Hispanic Asian (only)	1.04%	7.95%	5.86%	226,062
5 Non-Hispanic NHOPI (only)	2.19%	14.07%	0.23%	8,905
6 Non-Hispanic more than one race	1.59%	9.87%	1.88%	72,678
7 Hispanic	1.46%	9.97%	22.43%	865,385
8 Origin unknown or not stated	2.19%	10.12%	0.73%	28,075
Total Reported	1.52%	9.65%	96.63%	3,728,838
Missing Records (coded .)	1.45%	8.78%	3.37%	129,919
Total	1.52%	9.62%	100%	3,858,757

Appendix II - JMP Binary Logistic Regression Model Code

Final Binary Logistic Regression Model (Premature birth <37 weeks gestation)

```
(-1.19330982407569) + Match( :MotherRaceHisp 6,
	5, -0.0478422368424318,
	4, -0.0822859571496837,
	3, -0.0622803196382426,
	2, 0.262472175792844,
	6, -0.0417985472537607,
	1, -0.0282651149087253,
	.
) + Match( :BMI Levels,
	"Underweight", 0.0605356437117031,
	"Overweight", -0.048119680431734,
	"Obese", 0.0114905080842116,
	"Obese2", 0.020180680293232,
	"Normal", -0.0440871516574127,
	.
) + Match( :Name( "TermOfFirstPrenatalVisit" ), 3, -0.410907021388604, 2,
0.133642839631135, 1, 0.27726418175747, . )
+Match( :Name( "NumOfPrenatalVisits (Nom)" ),
	1, 0.96228972726752,
	2, 0.270548792652301,
	3, -0.180044107044727,
	4, -0.567239926757253,
	5, -0.485554486117841,
	.
) + Match( :GestationalHypertension, 0, -0.239078979848329, 1, 0.239078979848329, . )
+ Match( :Hypertension Eclampsia,
	0, -0.522090547963129,
	1, 0.522090547963129,
	.
) + Match( :NoRiskFactorsDetermined, "N", 0.271906933300773, "Y", -0.271906933300773,
. ) + Match( :PaymentMethod,
	"Medicaid", 0.0727470178213298,
	"Not Reported", 0.102547913867973,
	"Other Payment", -0.0749643313743903,
	"Self Pay", -0.177147421575449,
	"Private Ins", 0.0768168212605366,
	.
) + Match( :sex, "F", -0.063434661015323, "M", 0.063434661015323, . ) + Match( :Name(
"TermOfFirstPrenatalVisit" ),
	3,
		Match( :Name( "NumOfPrenatalVisits" ),
			1, -0.404603553393318,
			2, -0.412729834158802,
			3, -0.0300486304885534,
			4, 0.476761092097738,
			5, 0.370620925942935,
			.
		),
	2,
		Match( :Name( "NumOfPrenatalVisits" ),
			1, 0.0468643037944028,
```

```
                2, -0.0458073325592771,
                3, -0.075238792303704,
                4, -0.0376253567566833,
                5, 0.111807177825262,
                .
        ),
    1,
        Match( :Name( "NumOfPrenatalVisits" ),
                1, 0.357739249598915,
                2, 0.458537166718079,
                3, 0.105287422792257,
                4, -0.439135735341054,
                5, -0.482428103768197,
                .
        ),
    .
) + Match( :Mother Education Groups,
    "Bachelor's and Above", -0.0932546337909648,
    "Some College or Associates Degree", 0.0129602121455699,
    "High School Grad or GED", 0.080294421645395,
    .
) + Match( :Mothers_Age_Groups3, 2, 0.334013462378098, 1, -0.180407914557997, 0, -
0.153605547820101, . ) + Match( :Mothers_Age_Groups3,
    2,
        Match( :BMI Levels,
                "Underweight", -0.131704499286726,
                "Overweight", 0.039440961835379,
                "Obese", 0.0872010367796663,
                "Obese2", 0.051048441100982,
                "Normal", -0.0459859404293015,
                .
        ),
    1,
        Match( :BMI Levels,
                "Underweight", -0.0541100869571328,
                "Overweight", -0.00646225235935679,
                "Obese", 0.0155241211126896,
                "Obese2", 0.0998275526012867,
                "Normal", -0.0547793343974867,
                .
        ),
    0,
        Match( :BMI Levels,
                "Underweight", 0.185814586243859,
                "Overweight", -0.0329787094760222,
                "Obese", -0.102725157892356,
                "Obese2", -0.150875993702269,
                "Normal", 0.100765274826788,
                .
        ),
    .
)
```

Appendix III – SAS Binary Logistic Regression Model Code

```
Libname Data 'c:\SAS\Data\';


PROC IMPORT OUT= data.Natalityfile
            DATAFILE= "c:\sas\Final Thesis Dataset 5.2.csv"
            DBMS=csv REPLACE;
     GETNAMES=YES;
     DATAROW=2;
       guessingrows=10000;
 RUN;

data TestDataset (Where = (Var82 = 'Training'));
set Natalityfile;
run;

data ValidationDataset (Where = (Var82 = 'Validation'));
set Natalityfile;
run;

proc logistic data = TestDataset outmodel=OutModel;
Class NumOfPrenatalVisits (ref = '11-16 Visits') TermOfFirstPrenatalVisit
(ref = 'Second Trimester Visit') BMI_Levels (ref='Normal')
MotherAgeGroup3 (ref = '20-40') sex (ref ='M') Mother_Education_Nom (ref =
'HighSchool Grad or GED') MotherRaceHisp_6 (ref = 'Non-Hispanic White')
NoRiskFactorsDetermined (ref ='Y')PaymentMethod (ref = 'Private Ins')
GestationalHypertension (ref = 'N') Hypertension_Eclampsia (ref = 'N');
model Premature___37_wks_ (event='Premature <37 Weeks')= PaymentMethod Sex
GestationalHypertension Hypertension_Eclampsia NoRiskFactorsDetermined
Mother_Education_Nom
MotherRaceHisp_6 NumOfPrenatalVisits TermOfFirstPrenatalVisit
NumOfPrenatalVisits*TermOfFirstPrenatalVisit BMI_Levels*MotherAgeGroup3
BMI_Levels MotherAgeGroup3 / ctable pprob=0.5;
oddsratio NumOfPrenatalVisits;
oddsratio BMI_Levels;
oddsratio MotherAgeGroup3;
oddsratio sex;
oddsratio PaymentMethod;
oddsratio GestationalHypertension;
oddsratio Hypertension_Eclampsia;
oddsratio NoRiskFactorsDetermined;
oddsratio Mother_Education_Nom;
oddsratio MotherRaceHisp_6;
oddsratio TermOfFirstPrenatalVisit;
roc; roccontrast;
output out=preds predprobs=individual prob=p resdev=dr h=pii reschi=pr
difchisq=difchi ;
ods output ParameterEstimates;
score data=ValidationDataset fitstat ;
score data=TestDataset fitstat;
run;
```

Appendix VI – JMP Binary Logistic Regression Model Output

Final Thesis Dataset 5.2 - Fit Nominal Logistic

Nominal Logistic Fit for Premature (<37 wks)

Effect Summary

Source	LogWorth	PValue
NumOfPrenatalVisits (6 levels Nom)	1331.245	0.00000
NoRiskFactorsDetermined	368.854	0.00000
TermOfFirstPrenatalVisit (4 Groups)*NumOfPrenatalVisits (6 levels Nom)	330.416	0.00000
GestationalHypertension	201.043	0.00000
MotherRaceHisp 6	176.030	0.00000
Hypertension Eclampsia	114.442	0.00000
TermOfFirstPrenatalVisit (4 Groups)	68.970	0.00000 ^
sex	61.927	0.00000
Mother Education Nom	48.938	0.00000
PaymentMethod	39.791	0.00000
BMI Levels*MotherAgeGroup3	33.929	0.00000
MotherAgeGroup3	25.250	0.00000 ^
BMI Levels	1.451	0.03537 ^

Converged in Gradient, 6 iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	25944.58	43	51889.16	<.0001*
Full	252144.59			
Reduced	278089.17			

RSquare (U)	0.0933
AICc	504377
BIC	504893
Observations (or Sum Wgts)	920341

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	28573	15167.87	30335.73
Saturated	28616	236976.73	Prob>ChiSq
Fitted	43	252144.59	<.0001*

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-1.1933098	0.0348013	1175.8	<.0001*
TermOfFirstPrenatalVisit (4 Groups)[3rd Trimester Visit]	-0.410907	0.0386108	113.26	<.0001*
TermOfFirstPrenatalVisit (4 Groups)[Second Trimester Visit]	0.13364284	0.0213128	39.32	<.0001*
MotherRaceHisp 6[Hispanic]	-0.0417985	0.0109292	14.63	0.0001*
MotherRaceHisp 6[Non-Hispanic Two or More Races]	-0.0478422	0.0228023	4.40	0.0359*
MotherRaceHisp 6[Non-Hispanic Asian]	-0.082286	0.0156993	27.47	<.0001*
MotherRaceHisp 6[Non-Hispanic AIAN & NHOPI]	-0.0622803	0.0323367	3.71	0.0541
MotherRaceHisp 6[Non-Hispanic Black]	0.26247218	0.0113104	538.53	<.0001*
NumOfPrenatalVisits (6 levels Nom)[1-6 Visits]	0.96228973	0.021252	2050.3	<.0001*
NumOfPrenatalVisits (6 levels Nom)[7-8 Visits]	0.27054879	0.0238213	128.99	<.0001*
NumOfPrenatalVisits (6 levels Nom)[9-10 Visits]	-0.1800441	0.0265929	45.84	<.0001*
NumOfPrenatalVisits (6 levels Nom)[17+ Visits]	-0.4855545	0.0722815	45.13	<.0001*
NoRiskFactorsDetermined[N]	0.27190693	0.0063386	1840.2	<.0001*
MotherAgeGroup3[Over 40]	0.33401346	0.027839	143.95	<.0001*
MotherAgeGroup3[Teenager]	-0.1536055	0.0166091	85.53	<.0001*
GestationalHypertension[N]	-0.239079	0.0079151	912.36	<.0001*
Hypertension Eclampsia[N]	-0.5220905	0.0217617	575.58	<.0001*
BMI Levels[Underweight]	0.06053564	0.0465147	1.69	0.1931
BMI Levels[Overweight]	-0.0481197	0.0197067	5.96	0.0146*
BMI Levels[Obese]	0.01149051	0.0239332	0.23	0.6312
BMI Levels[VObese]	0.02018068	0.0260705	0.60	0.4389
BMI Levels[Underweight]*MotherAgeGroup3[Over 40]	-0.1317045	0.091226	2.08	0.1488
BMI Levels[Underweight]*MotherAgeGroup3[Teenager]	0.18581459	0.0489068	14.44	0.0001*
BMI Levels[Overweight]*MotherAgeGroup3[Over 40]	0.03944096	0.037472	1.11	0.2925
BMI Levels[Overweight]*MotherAgeGroup3[Teenager]	-0.0329787	0.0226542	2.12	0.1455
BMI Levels[Obese]*MotherAgeGroup3[Over 40]	0.08720104	0.0448356	3.78	0.0518
BMI Levels[Obese]*MotherAgeGroup3[Teenager]	-0.1027252	0.0285053	12.99	0.0003*
BMI Levels[VObese]*MotherAgeGroup3[Over 40]	0.05104844	0.0479983	1.13	0.2875
BMI Levels[VObese]*MotherAgeGroup3[Teenager]	-0.150876	0.0321932	21.96	<.0001*
Mother Education Nom[HighSchool Grad or GED]	0.08029442	0.0063083	162.01	<.0001*
Mother Education Nom[Bachelor's and Above]	-0.0932546	0.0066161	198.67	<.0001*
sex[F]	-0.0634347	0.0038012	278.49	<.0001*
TermOfFirstPrenatalVisit (4 Groups)[3rd Trimester Visit]*NumOfPrenatalVisits (6 levels Nom)[1-6 Visits]	-0.4046036	0.0404482	100.06	<.0001*
TermOfFirstPrenatalVisit (4 Groups)[3rd Trimester Visit]*NumOfPrenatalVisits (6 levels Nom)[7-8 Visits]	-0.4127298	0.04591	80.82	<.0001*
TermOfFirstPrenatalVisit (4 Groups)[3rd Trimester Visit]*NumOfPrenatalVisits (6 levels Nom)[9-10 Visits]	-0.0300486	0.0518111	0.34	0.5619
TermOfFirstPrenatalVisit (4 Groups)[3rd Trimester Visit]*NumOfPrenatalVisits (6 levels Nom)[17+ Visits]	0.37062093	0.1415271	6.86	0.0088*
TermOfFirstPrenatalVisit (4 Groups)[Second Trimester Visit]*NumOfPrenatalVisits (6 levels Nom)[1-6 Visits]	0.0468643	0.0237308	3.90	0.0483*
TermOfFirstPrenatalVisit (4 Groups)[Second Trimester Visit]*NumOfPrenatalVisits (6 levels Nom)[7-8 Visits]	-0.0458073	0.0262682	3.04	0.0812
TermOfFirstPrenatalVisit (4 Groups)[Second Trimester Visit]*NumOfPrenatalVisits (6 levels Nom)[9-10 Visits]	-0.0752388	0.0287407	6.85	0.0088*
TermOfFirstPrenatalVisit (4 Groups)[Second Trimester Visit]*NumOfPrenatalVisits (6 levels Nom)[17+ Visits]	0.11180718	0.0774461	2.08	0.1488
PaymentMethod[Medicaid]	0.07274702	0.0116497	38.99	<.0001*
PaymentMethod[Not Reported]	0.10254791	0.0349399	8.61	0.0033*
PaymentMethod[Other Payment]	-0.0749643	0.0179447	17.45	<.0001*
PaymentMethod[Self Pay]	-0.1771474	0.0198247	79.85	<.0001*

For log odds of Premature <37 Weeks/Normal Birth >=37 weeks

Nominal Logistic Fit for Premature (<37 wks)

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
TermOfFirstPrenatalVisit (4 Groups)	2	2	317.617102	<.0001*
MotherRaceHisp 6	5	5	828.162478	<.0001*
NumOfPrenatalVisits (6 levels Nom)	4	4	6146.67054	<.0001*
NoRiskFactorsDetermined	1	1	1690.74827	<.0001*
MotherAgeGroup3	2	2	116.281687	<.0001*
GestationalHypertension	1	1	918.562922	<.0001*
Hypertension Eclampsia	1	1	520.312766	<.0001*
BMI Levels	4	4	10.3198678	0.0354*
BMI Levels*MotherAgeGroup3	8	8	179.720844	<.0001*
Mother Education Nom	2	2	225.367184	<.0001*
sex	1	1	279.095042	<.0001*
TermOfFirstPrenatalVisit (4 Groups)*NumOfPrenatalVisits (6 levels Nom)	8	8	1557.99261	<.0001*
PaymentMethod	4	4	192.397206	<.0001*

Odds Ratios

For Premature (<37 wks) odds of Premature <37 Weeks versus
Normal Birth >=37 weeks

Odds Ratios for TermOfFirstPrenatalVisit (4 Groups)

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
Second Trimester Visit	3rd Trimester Visit	1.7238322	<.0001*	1.5354354	1.9353452
First Trimester Visit	3rd Trimester Visit	1.9900728	<.0001*	1.7771434	2.2285144
First Trimester Visit	Second Trimester Visit	1.1544469	<.0001*	1.1210752	1.188812
3rd Trimester Visit	Second Trimester Visit	0.5801029	<.0001*	0.5167037	0.6512811
3rd Trimester Visit	First Trimester Visit	0.5024942	<.0001*	0.4487294	0.5627008
Second Trimester Visit	First Trimester Visit	0.8662157	<.0001*	0.8411759	0.8920008

Odds Ratios for MotherRaceHisp 6

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
Non-Hispanic Two or More Races	Hispanic	0.9939745	0.8248	0.9421894	1.0486059
Non-Hispanic Asian	Hispanic	0.9603213	0.0273*	0.926407	0.995477
Non-Hispanic Asian	Non-Hispanic Two or More Races	0.9661427	0.2596	0.9099862	1.0257647
Non-Hispanic AIAN & NHOPI	Hispanic	0.9797266	0.6001	0.9075085	1.0576915
Non-Hispanic AIAN & NHOPI	Non-Hispanic Two or More Races	0.9856656	0.7544	0.9004076	1.0789967
Non-Hispanic AIAN & NHOPI	Non-Hispanic Asian	1.0202071	0.6308	0.9402687	1.1069415
Non-Hispanic Black	Hispanic	1.355636	<.0001*	1.3231129	1.3889586
Non-Hispanic Black	Non-Hispanic Two or More Races	1.3638539	<.0001*	1.2922313	1.4394461
Non-Hispanic Black	Non-Hispanic Asian	1.4116484	<.0001*	1.3607823	1.464416
Non-Hispanic Black	Non-Hispanic AIAN & NHOPI	1.3836881	<.0001*	1.281325	1.4942289
Non-Hispanic White	Hispanic	1.0136254	0.1913	0.9932586	1.0344099
Non-Hispanic White	Non-Hispanic Two or More Races	1.01977	0.4609	0.9680611	1.074241
Non-Hispanic White	Non-Hispanic Asian	1.0555066	0.0012*	1.0216196	1.0905177
Non-Hispanic White	Non-Hispanic AIAN & NHOPI	1.0346003	0.3795	0.9590131	1.1161452
Non-Hispanic White	Non-Hispanic Black	0.7477121	<.0001*	0.7318099	0.7639598
Hispanic	Non-Hispanic Two or More Races	1.006062	0.8248	0.9536471	1.0613577
Hispanic	Non-Hispanic Asian	1.0413182	0.0273*	1.0045435	1.0794391
Non-Hispanic Two or More Races	Non-Hispanic Asian	1.0350438	0.2596	0.9748824	1.0989178
Hispanic	Non-Hispanic AIAN & NHOPI	1.020693	0.6001	0.9454552	1.101918
Non-Hispanic Two or More Races	Non-Hispanic AIAN & NHOPI	1.0145428	0.7544	0.9267869	1.1106082
Non-Hispanic Asian	Non-Hispanic AIAN & NHOPI	0.9801931	0.6308	0.9033901	1.0635257
Hispanic	Non-Hispanic Black	0.7376611	<.0001*	0.7199639	0.7557934
Non-Hispanic Two or More Races	Non-Hispanic Black	0.7332164	<.0001*	0.6947117	0.7738553
Non-Hispanic Asian	Non-Hispanic Black	0.7083917	<.0001*	0.6828661	0.7348714
Non-Hispanic AIAN & NHOPI	Non-Hispanic Black	0.7227062	<.0001*	0.6692415	0.7804421
Hispanic	Non-Hispanic White	0.9865577	0.1913	0.9667348	1.0067872
Non-Hispanic Two or More Races	Non-Hispanic White	0.9806133	0.4609	0.9308898	1.0329927
Non-Hispanic Asian	Non-Hispanic White	0.9474124	0.0012*	0.9169957	0.9788379
Non-Hispanic AIAN & NHOPI	Non-Hispanic White	0.9665568	0.3795	0.8959408	1.0427386
Non-Hispanic Black	Non-Hispanic White	1.3374132	<.0001*	1.3089694	1.366475

Odds Ratios for NumOfPrenatalVisits (6 levels Nom)

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
7-8 Visits	1-6 Visits	0.5007036	<.0001*	0.4816235	0.5205396
9-10 Visits	1-6 Visits	0.3190735	<.0001*	0.3037929	0.3351227
9-10 Visits	7-8 Visits	0.6372502	<.0001*	0.6024579	0.6740519
17+ Visits	1-6 Visits	0.2350765	<.0001*	0.1969372	0.280602
17+ Visits	7-8 Visits	0.4694923	<.0001*	0.3925048	0.5615806
17+ Visits	9-10 Visits	0.7367473	0.0010*	0.6144148	0.8834365
11-16 Visits	1-6 Visits	0.2166375	<.0001*	0.2042426	0.2297847
11-16 Visits	7-8 Visits	0.4326662	<.0001*	0.4054778	0.4616777
11-16 Visits	9-10 Visits	0.6789581	<.0001*	0.632137	0.7292472
11-16 Visits	17+ Visits	0.9215618	0.3854	0.766344	1.1082179
1-6 Visits	7-8 Visits	1.9971895	<.0001*	1.9210834	2.0763106
1-6 Visits	9-10 Visits	3.1340742	<.0001*	2.9839817	3.2917164
7-8 Visits	9-10 Visits	1.5692423	<.0001*	1.4835654	1.6598672
1-6 Visits	17+ Visits	4.253934	<.0001*	3.5637665	5.0777611
7-8 Visits	17+ Visits	2.1299602	<.0001*	1.7806884	2.5477396
9-10 Visits	17+ Visits	1.3573176	0.0010*	1.1319432	1.6275648
1-6 Visits	11-16 Visits	4.6160052	<.0001*	4.3518993	4.896139
7-8 Visits	11-16 Visits	2.3112505	<.0001*	2.1660132	2.4662264
9-10 Visits	11-16 Visits	1.4728449	<.0001*	1.3712771	1.5819355
17+ Visits	11-16 Visits	1.0851144	0.3854	0.9023496	1.304897

Nominal Logistic Fit for Premature (<37 wks)

Odds Ratios

Odds Ratios for NoRiskFactorsDetermined

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
Y	N	0.58053	<.0001*	0.5662834	0.5951349
N	Y	1.722564	<.0001*	1.6802912	1.7659002

Odds Ratios for MotherAgeGroup3

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
Teenager	Over 40	0.6140868	<.0001*	0.5640057	0.6686148
20-40	Over 40	0.5978464	<.0001*	0.5513074	0.6483141
20-40	Teenager	0.9735536	0.0658	0.9461474	1.0017537
Over 40	Teenager	1.6284343	<.0001*	1.4956294	1.7730317
Over 40	20-40	1.6726704	<.0001*	1.5424622	1.8138702
Teenager	20-40	1.0271648	0.0658	0.9982494	1.0569177

Odds Ratios for GestationalHypertension

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
Y	N	1.6131003	<.0001*	1.5638195	1.663934
N	Y	0.6199243	<.0001*	0.6009854	0.63946

Odds Ratios for Hypertension Eclampsia

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
Y	N	2.841071	<.0001*	2.6087648	3.0940637
N	Y	0.3519799	<.0001*	0.3231995	0.3833232

Odds Ratios for BMI Levels

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
Overweight	Underweight	0.8970396	0.0688	0.7979807	1.0083952
Obese	Underweight	0.9521382	0.4308	0.8427798	1.0756868
Obese	Overweight	1.0614227	0.0487*	1.0003504	1.1262236
VObese	Underweight	0.9604485	0.5262	0.8477643	1.0881104
VObese	Overweight	1.0706869	0.0386*	1.0035815	1.1422793
VObese	Obese	1.008728	0.8159	0.9375664	1.0852909
Normal	Underweight	0.9006642	0.0751	0.8026554	1.0106405
Normal	Overweight	1.0040407	0.8573	0.9608519	1.0491707
Normal	Obese	0.9459386	0.0512	0.8945449	1.0002848
Normal	VObese	0.9377538	0.0412*	0.8816469	0.9974312
Underweight	Overweight	1.114778	0.0688	0.9916747	1.2531631
Underweight	Obese	1.0502678	0.4308	0.9296386	1.1865496
Overweight	Obese	0.9421317	0.0487*	0.8879232	0.9996497
Underweight	VObese	1.0411803	0.5262	0.9190244	1.1795731
Overweight	VObese	0.9339799	0.0386*	0.8754427	0.9964313
Obese	VObese	0.9913475	0.8159	0.921412	1.0665911
Underweight	Normal	1.1102917	0.0751	0.9894716	1.2458647
Overweight	Normal	0.9959756	0.8573	0.9531338	1.0407431
Obese	Normal	1.0571511	0.0512	0.9997152	1.1178868
VObese	Normal	1.066378	0.0412*	1.0025754	1.1342409

Odds Ratios for Mother Education Nom

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
Bachelor's and Above	HighSchool Grad or GED	0.8406759	<.0001*	0.8217617	0.8600255
Some College or Associates Degee	HighSchool Grad or GED	0.9348827	<.0001*	0.9166937	0.9534327
Some College or Associates Degee	Bachelor's and Above	1.1120608	<.0001*	1.0891888	1.1354131
HighSchool Grad or GED	Bachelor's and Above	1.189519	<.0001*	1.1627562	1.2168978
HighSchool Grad or GED	Some College or Associates Degee	1.0696529	<.0001*	1.0488418	1.090877
Bachelor's and Above	Some College or Associates Degee	0.8992314	<.0001*	0.8807367	0.9181145

Odds Ratios for sex

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
M	F	1.1352687	<.0001*	1.1184781	1.1523113
F	M	0.8808488	<.0001*	0.8678211	0.894072

Odds Ratios for PaymentMethod

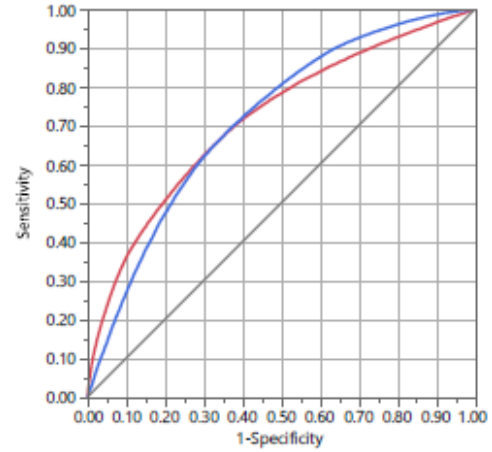
Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
Not Reported	Medicaid	1.0302494	0.4934	0.9460404	1.121954
Other Payment	Medicaid	0.8626801	<.0001*	0.8298143	0.8968476
Other Payment	Not Reported	0.8373507	0.0002*	0.7637056	0.9180976
Self Pay	Medicaid	0.778883	<.0001*	0.7452101	0.8140774
Self Pay	Not Reported	0.756014	<.0001*	0.6878629	0.8309174
Self Pay	Other Payment	0.9028642	0.0004*	0.8534467	0.9551432
Private Ins	Medicaid	1.0040781	0.6747	0.9851729	1.0233461
Private Ins	Not Reported	0.9745971	0.5541	0.89497	1.0613088
Private Ins	Other Payment	1.1639055	<.0001*	1.1193982	1.2101824
Private Ins	Self Pay	1.2891257	<.0001*	1.2328806	1.3479367
Medicaid	Not Reported	0.9706388	0.4934	0.8913021	1.0570373
Medicaid	Other Payment	1.1591783	<.0001*	1.1150167	1.2050889
Not Reported	Other Payment	1.1942427	0.0002*	1.0892088	1.3094051
Medicaid	Self Pay	1.2838899	<.0001*	1.2283844	1.3419034
Not Reported	Self Pay	1.3227268	<.0001*	1.2034891	1.4537781
Other Payment	Self Pay	1.1075862	0.0004*	1.0469635	1.1717193
Medicaid	Private Ins	0.9959385	0.6747	0.9771866	1.0150502
Not Reported	Private Ins	1.026065	0.5541	0.9422328	1.1173558
Other Payment	Private Ins	0.8591763	<.0001*	0.8263217	0.8933371
Self Pay	Private Ins	0.7757195	<.0001*	0.7418746	0.8111085

Nominal Logistic Fit for Premature (<37 wks)

Odds Ratios

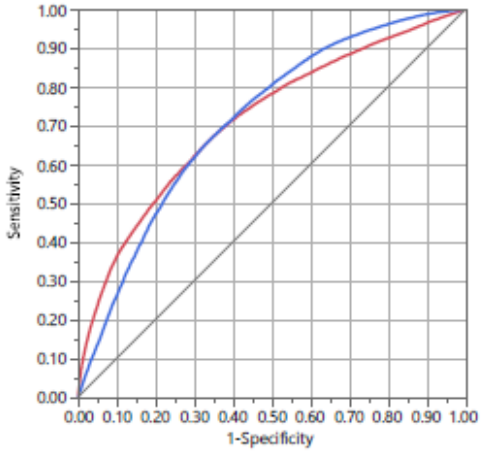
Normal approximations used for ratio confidence limits effects:
TermOfFirstPrenatalVisit (4 Groups) MotherRaceHisp 6
NumOfPrenatalVisits (6 levels Nom) NoRiskFactorsDetermined
MotherAgeGroup3 GestationalHypertension Hypertension
Eclampsia BMI Levels Mother Education Nom sex PaymentMethod
Tests and confidence intervals on odds ratios are Wald based.

Receiver Operating Characteristic



Premature (<37 wks)	Area
Premature <37 Weeks	0.7166
Normal Birth >=37 weeks	0.7166

Receiver Operating Characteristic on Validation Data



Premature (<37 wks)	Area
Premature <37 Weeks	0.7156
Normal Birth >=37 weeks	0.7156

Confusion Matrix

Training

Actual Premature (<37 wks)	Predicted Count Premature <37 Weeks	Predicted Count Normal Birth >=37 weeks
Premature <37 Weeks	2004	80676
Normal Birth >=37 weeks	1598	836063

Validation

Actual Premature (<37 wks)	Predicted Count Premature <37 Weeks	Predicted Count Normal Birth >=37 weeks
Premature <37 Weeks	833	34441
Normal Birth >=37 weeks	671	357959

Appendix V – SAS Binary Logistic Regression Model Output

The SAS System

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TESTDATASET
Response Variable	Premature___37_wks_
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	920341
Number of Observations Used	920341

Response Profile		
Ordered Value	Premature___37_wks_	Total Frequency
1	Normal Birth >=37 weeks	837661
2	Premature <37 Weeks	82680

Probability modeled is Premature___37_wks_='Premature <37 Weeks'.

Class Level Information						
Class	Value	Design Variables				
NumOfPrenatalVisits	1-6 Visits	1	0	0	0	
	11-16 Visits	-1	-1	-1	-1	
	17+ Visits	0	1	0	0	
	7-8 Visits	0	0	1	0	
	9-10 Visits	0	0	0	1	
TermOfFirstPrenatalVisit	3rd Trimester Visit	1	0			
	First Trimester Visit	0	1			
	Second Trimester Visit	-1	-1			
BMI_Levels	Normal	-1	-1	-1	-1	
	ObeseClass1	1	0	0	0	
	ObeseClass2	0	1	0	0	
	Overweight	0	0	1	0	
	Underweight	0	0	0	1	
MotherAgeGroup3	20-40	-1	-1			
	Over 40	1	0			
	Teenager	0	1			
sex	F	1				
	M	-1				
Mother_Education_Nom	Bachelor's and Above	1	0			
	HighSchool Grad or GED	-1	-1			
	Some College or Associates Degee	0	1			
MotherRaceHisp_6	Hispanic	1	0	0	0	0
	Non-Hispanic AIAN & NHOPI	0	1	0	0	0
	Non-Hispanic Asian	0	0	1	0	0
	Non-Hispanic Black	0	0	0	1	0
	Non-Hispanic Two or More Races	0	0	0	0	1
	Non-Hispanic White	-1	-1	-1	-1	-1
NoRiskFactorsDetermined	N	1				
	Y	-1				

PaymentMethod	Medicaid	1	0	0	0	
	Not Reported	0	1	0	0	
	Other Payment	0	0	1	0	
	Private Ins	-1	-1	-1	-1	
	Self Pay	0	0	0	1	
GestationalHypertension	N	-1				
	Y	1				
Hypertension_Eclampsia	N	-1				
	Y	1				

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	556180.34	504377.18
SC	556192.07	504893.41
-2 Log L	556178.34	504289.18

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	51889.1572	43	<.0001
Score	65319.7516	43	<.0001
Wald	52130.0613	43	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
PaymentMethod	4	183.9831	<.0001
sex	1	278.4912	<.0001
GestationalHypertens	1	912.3599	<.0001
Hypertension_Eclamps	1	575.5787	<.0001
NoRiskFactorsDetermi	1	1840.1606	<.0001
Mother_Education_Nom	2	225.3434	<.0001
MotherRaceHisp_6	5	859.1633	<.0001
NumOfPrenatalVisits	4	4656.2747	<.0001
TermOfFirstPrenatalV	2	225.1715	<.0001
NumOfPren*TermOfFirs	8	1740.4316	<.0001
BMI_Level*MotherAgeG	8	175.9676	<.0001
BMI_Levels	4	10.5135	0.0326
MotherAgeGroup3	2	156.8792	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter			DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept			1	-1.1933	0.0348	1175.7529	<.0001
PaymentMethod	Medicaid		1	0.0727	0.0116	38.9936	<.0001
PaymentMethod	Not Reported		1	0.1025	0.0349	8.6140	0.0033
PaymentMethod	Other Payment		1	-0.0750	0.0179	17.4516	<.0001
PaymentMethod	Self Pay		1	-0.1771	0.0198	79.8452	<.0001
sex	F		1	-0.0634	0.00380	278.4912	<.0001
GestationalHypertens	Y		1	0.2391	0.00792	912.3599	<.0001
Hypertension_Eclamps	Y		1	0.5221	0.0218	575.5787	<.0001

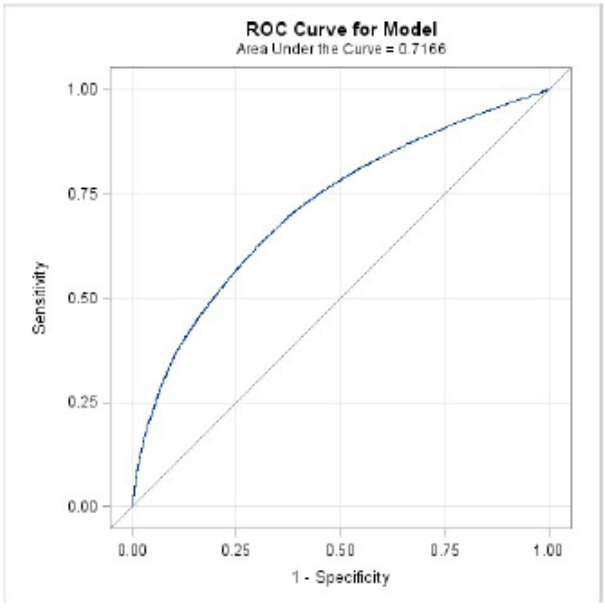
NoRiskFactorsDetermi	N		1	0.2719	0.00634	1840.1606	<.0001
Mother_Education_Nom	Bachelor's and Above		1	-0.0933	0.00662	198.6692	<.0001
Mother_Education_Nom	Some College or Associates Degee		1	0.0130	0.00569	5.1942	0.0227
MotherRaceHisp_6	Hispanic		1	-0.0418	0.0109	14.6268	0.0001
MotherRaceHisp_6	Non-Hispanic AIAN & NHOPI		1	-0.0623	0.0323	3.7094	0.0541
MotherRaceHisp_6	Non-Hispanic Asian		1	-0.0823	0.0157	27.4718	<.0001
MotherRaceHisp_6	Non-Hispanic Black		1	0.2625	0.0113	538.5305	<.0001
MotherRaceHisp_6	Non-Hispanic Two or More Races		1	-0.0478	0.0228	4.4021	0.0359
NumOfPrenatalVisits	1-6 Visits		1	0.9623	0.0213	2050.2929	<.0001
NumOfPrenatalVisits	17+ Visits		1	-0.4856	0.0723	45.1256	<.0001
NumOfPrenatalVisits	7-8 Visits		1	0.2705	0.0238	128.9899	<.0001
NumOfPrenatalVisits	9-10 Visits		1	-0.1800	0.0266	45.8387	<.0001
TermOfFirstPrenatalV	3rd Trimester Visit		1	-0.4109	0.0386	113.2590	<.0001
TermOfFirstPrenatalV	First Trimester Visit		1	0.2773	0.0201	190.6890	<.0001
NumOfPren*TermOfFirs	1-6 Visits	3rd Trimester Visit	1	-0.4046	0.0404	100.0624	<.0001
NumOfPren*TermOfFirs	1-6 Visits	First Trimester Visit	1	0.3577	0.0224	254.8810	<.0001
NumOfPren*TermOfFirs	17+ Visits	3rd Trimester Visit	1	0.3706	0.1415	6.8578	0.0088
NumOfPren*TermOfFirs	17+ Visits	First Trimester Visit	1	-0.4824	0.0728	43.9148	<.0001
NumOfPren*TermOfFirs	7-8 Visits	3rd Trimester Visit	1	-0.4127	0.0459	80.8191	<.0001
NumOfPren*TermOfFirs	7-8 Visits	First Trimester Visit	1	0.4585	0.0246	347.9228	<.0001
NumOfPren*TermOfFirs	9-10 Visits	3rd Trimester Visit	1	-0.0300	0.0518	0.3363	0.5620
NumOfPren*TermOfFirs	9-10 Visits	First Trimester Visit	1	0.1053	0.0270	15.1539	<.0001
BMI_Level*MotherAgeG	ObeseClass1	Over 40	1	0.0872	0.0448	3.7827	0.0518
BMI_Level*MotherAgeG	ObeseClass1	Teenager	1	-0.1027	0.0285	12.9868	0.0003
BMI_Level*MotherAgeG	ObeseClass2	Over 40	1	0.0510	0.0480	1.1311	0.2875
BMI_Level*MotherAgeG	ObeseClass2	Teenager	1	-0.1509	0.0322	21.9640	<.0001
BMI_Level*MotherAgeG	Overweight	Over 40	1	0.0394	0.0375	1.1078	0.2926
BMI_Level*MotherAgeG	Overweight	Teenager	1	-0.0330	0.0227	2.1192	0.1455
BMI_Level*MotherAgeG	Underweight	Over 40	1	-0.1317	0.0912	2.0843	0.1488
BMI_Level*MotherAgeG	Underweight	Teenager	1	0.1858	0.0489	14.4351	0.0001
BMI_Levels	ObeseClass1		1	0.0115	0.0239	0.2305	0.6311
BMI_Levels	ObeseClass2		1	0.0202	0.0261	0.5992	0.4389
BMI_Levels	Overweight		1	-0.0481	0.0197	5.9623	0.0146
BMI_Levels	Underweight		1	0.0605	0.0465	1.6937	0.1931
MotherAgeGroup3	Over 40		1	0.3340	0.0278	143.9525	<.0001
MotherAgeGroup3	Teenager		1	-0.1536	0.0166	85.5300	<.0001

Odds Ratio Estimates and Wald Confidence Intervals			
Label	Estimate	95% Confidence Limits	
NumOfPrenatalVisits 1-6 Visits vs 11-16 Visits at TermOfFirstPrenatalVisit=3rd Trimester Visit	1.912	1.617	2.261
NumOfPrenatalVisits 1-6 Visits vs 17+ Visits at TermOfFirstPrenatalVisit=3rd Trimester Visit	1.959	1.171	3.279
NumOfPrenatalVisits 1-6 Visits vs 7-8 Visits at TermOfFirstPrenatalVisit=3rd Trimester Visit	2.013	1.821	2.227
NumOfPrenatalVisits 1-6 Visits vs 9-10 Visits at TermOfFirstPrenatalVisit=3rd Trimester Visit	2.155	1.881	2.469
NumOfPrenatalVisits 17+ Visits vs 11-16 Visits at TermOfFirstPrenatalVisit=3rd Trimester Visit	0.976	0.570	1.672
NumOfPrenatalVisits 7-8 Visits vs 11-16 Visits at TermOfFirstPrenatalVisit=3rd Trimester Visit	0.950	0.788	1.145
NumOfPrenatalVisits 9-10 Visits vs 11-16 Visits at TermOfFirstPrenatalVisit=3rd Trimester Visit	0.887	0.721	1.092
NumOfPrenatalVisits 17+ Visits vs 7-8 Visits at TermOfFirstPrenatalVisit=3rd Trimester Visit	1.028	0.610	1.731
NumOfPrenatalVisits 17+ Visits vs 9-10 Visits at TermOfFirstPrenatalVisit=3rd Trimester Visit	1.100	0.648	1.867
NumOfPrenatalVisits 7-8 Visits vs 9-10 Visits at TermOfFirstPrenatalVisit=3rd Trimester Visit	1.070	0.913	1.255
NumOfPrenatalVisits 1-6 Visits vs 11-16 Visits at TermOfFirstPrenatalVisit=First Trimester Visit	10.241	9.926	10.567
NumOfPrenatalVisits 1-6 Visits vs 17+ Visits at TermOfFirstPrenatalVisit=First Trimester Visit	9.855	9.411	10.321

NumOfPrenatalVisits 1-6 Visits vs 7-8 Visits at TermOfFirstPrenatalVisit=First Trimester Visit	1.806	1.742	1.872
NumOfPrenatalVisits 1-6 Visits vs 9-10 Visits at TermOfFirstPrenatalVisit=First Trimester Visit	4.034	3.905	4.168
NumOfPrenatalVisits 17+ Visits vs 11-16 Visits at TermOfFirstPrenatalVisit=First Trimester Visit	1.039	1.000	1.080
NumOfPrenatalVisits 7-8 Visits vs 11-16 Visits at TermOfFirstPrenatalVisit=First Trimester Visit	5.672	5.524	5.823
NumOfPrenatalVisits 9-10 Visits vs 11-16 Visits at TermOfFirstPrenatalVisit=First Trimester Visit	2.539	2.486	2.593
NumOfPrenatalVisits 17+ Visits vs 7-8 Visits at TermOfFirstPrenatalVisit=First Trimester Visit	0.183	0.176	0.191
NumOfPrenatalVisits 17+ Visits vs 9-10 Visits at TermOfFirstPrenatalVisit=First Trimester Visit	0.409	0.393	0.426
NumOfPrenatalVisits 7-8 Visits vs 9-10 Visits at TermOfFirstPrenatalVisit=First Trimester Visit	2.234	2.173	2.297
NumOfPrenatalVisits 1-6 Visits vs 11-16 Visits at TermOfFirstPrenatalVisit=Second Trimester Visit	5.023	4.797	5.260
NumOfPrenatalVisits 1-6 Visits vs 17+ Visits at TermOfFirstPrenatalVisit=Second Trimester Visit	3.986	3.531	4.501
NumOfPrenatalVisits 1-6 Visits vs 7-8 Visits at TermOfFirstPrenatalVisit=Second Trimester Visit	2.191	2.092	2.294
NumOfPrenatalVisits 1-6 Visits vs 9-10 Visits at TermOfFirstPrenatalVisit=Second Trimester Visit	3.541	3.384	3.705
NumOfPrenatalVisits 17+ Visits vs 11-16 Visits at TermOfFirstPrenatalVisit=Second Trimester Visit	1.260	1.115	1.423
NumOfPrenatalVisits 7-8 Visits vs 11-16 Visits at TermOfFirstPrenatalVisit=Second Trimester Visit	2.292	2.185	2.405
NumOfPrenatalVisits 9-10 Visits vs 11-16 Visits at TermOfFirstPrenatalVisit=Second Trimester Visit	1.418	1.353	1.487
NumOfPrenatalVisits 17+ Visits vs 7-8 Visits at TermOfFirstPrenatalVisit=Second Trimester Visit	0.550	0.486	0.621
NumOfPrenatalVisits 17+ Visits vs 9-10 Visits at TermOfFirstPrenatalVisit=Second Trimester Visit	0.888	0.786	1.003
NumOfPrenatalVisits 7-8 Visits vs 9-10 Visits at TermOfFirstPrenatalVisit=Second Trimester Visit	1.616	1.541	1.694
BMI_Levels ObeseClass1 vs Normal at MotherAgeGroup3=20-40	1.134	1.105	1.164
BMI_Levels ObeseClass2 vs Normal at MotherAgeGroup3=20-40	1.245	1.211	1.279
BMI_Levels Overweight vs Normal at MotherAgeGroup3=20-40	1.045	1.024	1.067
BMI_Levels Underweight vs Normal at MotherAgeGroup3=20-40	1.111	1.067	1.156
BMI_Levels ObeseClass1 vs ObeseClass2 at MotherAgeGroup3=20-40	0.911	0.882	0.942
BMI_Levels ObeseClass1 vs Overweight at MotherAgeGroup3=20-40	1.085	1.055	1.116
BMI_Levels ObeseClass1 vs Underweight at MotherAgeGroup3=20-40	1.021	0.976	1.067
BMI_Levels ObeseClass2 vs Overweight at MotherAgeGroup3=20-40	1.191	1.157	1.225
BMI_Levels ObeseClass2 vs Underweight at MotherAgeGroup3=20-40	1.120	1.071	1.172
BMI_Levels Overweight vs Underweight at MotherAgeGroup3=20-40	0.941	0.903	0.981
BMI_Levels ObeseClass1 vs Normal at MotherAgeGroup3=Over 40	1.208	1.038	1.406
BMI_Levels ObeseClass2 vs Normal at MotherAgeGroup3=Over 40	1.175	0.998	1.384
BMI_Levels Overweight vs Normal at MotherAgeGroup3=Over 40	1.085	0.960	1.226
BMI_Levels Underweight vs Normal at MotherAgeGroup3=Over 40	1.019	0.727	1.428
BMI_Levels ObeseClass1 vs ObeseClass2 at MotherAgeGroup3=Over 40	1.028	0.847	1.248
BMI_Levels ObeseClass1 vs Overweight at MotherAgeGroup3=Over 40	1.113	0.948	1.307
BMI_Levels ObeseClass1 vs Underweight at MotherAgeGroup3=Over 40	1.185	0.833	1.687
BMI_Levels ObeseClass2 vs Overweight at MotherAgeGroup3=Over 40	1.083	0.912	1.286
BMI_Levels ObeseClass2 vs Underweight at MotherAgeGroup3=Over 40	1.153	0.806	1.650
BMI_Levels Overweight vs Underweight at MotherAgeGroup3=Over 40	1.064	0.757	1.498
BMI_Levels ObeseClass1 vs Normal at MotherAgeGroup3=Teenager	0.863	0.808	0.921
BMI_Levels ObeseClass2 vs Normal at MotherAgeGroup3=Teenager	0.829	0.765	0.899
BMI_Levels Overweight vs Normal at MotherAgeGroup3=Teenager	0.871	0.833	0.912
BMI_Levels Underweight vs Normal at MotherAgeGroup3=Teenager	1.209	1.134	1.289
BMI_Levels ObeseClass1 vs ObeseClass2 at MotherAgeGroup3=Teenager	1.040	0.944	1.147
BMI_Levels ObeseClass1 vs Overweight at MotherAgeGroup3=Teenager	0.990	0.922	1.063
BMI_Levels ObeseClass1 vs Underweight at MotherAgeGroup3=Teenager	0.713	0.656	0.776
BMI_Levels ObeseClass2 vs Overweight at MotherAgeGroup3=Teenager	0.952	0.874	1.037
BMI_Levels ObeseClass2 vs Underweight at MotherAgeGroup3=Teenager	0.686	0.623	0.756
BMI_Levels Overweight vs Underweight at MotherAgeGroup3=Teenager	0.721	0.672	0.773
MotherAgeGroup3 Over 40 vs 20-40 at BMI_Levels=Normal	1.687	1.559	1.826
MotherAgeGroup3 Teenager vs 20-40 at BMI_Levels=Normal	1.200	1.164	1.237
MotherAgeGroup3 Over 40 vs Teenager at BMI_Levels=Normal	1.406	1.294	1.529

	1.797	1.574	2.051
MotherAgeGroup3 Teenager vs 20-40 at BMI_Levels=ObeseClass1	0.913	0.855	0.974
MotherAgeGroup3 Over 40 vs Teenager at BMI_Levels=ObeseClass1	1.969	1.705	2.274
MotherAgeGroup3 Over 40 vs 20-40 at BMI_Levels=ObeseClass2	1.593	1.377	1.843
MotherAgeGroup3 Teenager vs 20-40 at BMI_Levels=ObeseClass2	0.799	0.737	0.867
MotherAgeGroup3 Over 40 vs Teenager at BMI_Levels=ObeseClass2	1.993	1.692	2.347
MotherAgeGroup3 Over 40 vs 20-40 at BMI_Levels=Overweight	1.751	1.592	1.927
MotherAgeGroup3 Teenager vs 20-40 at BMI_Levels=Overweight	1.000	0.959	1.043
MotherAgeGroup3 Over 40 vs Teenager at BMI_Levels=Overweight	1.751	1.580	1.939
MotherAgeGroup3 Over 40 vs 20-40 at BMI_Levels=Underweight	1.548	1.112	2.154
MotherAgeGroup3 Teenager vs 20-40 at BMI_Levels=Underweight	1.306	1.217	1.401
MotherAgeGroup3 Over 40 vs Teenager at BMI_Levels=Underweight	1.185	0.849	1.655
sex F vs M	0.881	0.868	0.894
PaymentMethod Medicaid vs Not Reported	0.971	0.891	1.057
PaymentMethod Medicaid vs Other Payment	1.159	1.115	1.205
PaymentMethod Medicaid vs Private Ins	0.996	0.977	1.015
PaymentMethod Medicaid vs Self Pay	1.284	1.228	1.342
PaymentMethod Not Reported vs Other Payment	1.194	1.089	1.309
PaymentMethod Not Reported vs Private Ins	1.026	0.942	1.117
PaymentMethod Not Reported vs Self Pay	1.323	1.203	1.454
PaymentMethod Other Payment vs Private Ins	0.859	0.826	0.893
PaymentMethod Other Payment vs Self Pay	1.108	1.047	1.172
PaymentMethod Self Pay vs Private Ins	0.776	0.742	0.811
GestationalHypertension Y vs N	1.613	1.564	1.664
Hypertension_Eclampsia Y vs N	2.841	2.609	3.094
NoRiskFactorsDetermined N vs Y	1.723	1.680	1.766
Mother_Education_Nom Bachelor's and Above vs HighSchool Grad or GED	0.841	0.822	0.860
Mother_Education_Nom Bachelor's and Above vs Some College or Associates Degee	0.899	0.881	0.918
Mother_Education_Nom Some College or Associates Degee vs HighSchool Grad or GED	0.935	0.917	0.953
MotherRaceHisp_6 Hispanic vs Non-Hispanic AIAN & NHOPI	1.021	0.945	1.102
MotherRaceHisp_6 Hispanic vs Non-Hispanic Asian	1.041	1.005	1.079
MotherRaceHisp_6 Hispanic vs Non-Hispanic Black	0.738	0.720	0.756
MotherRaceHisp_6 Hispanic vs Non-Hispanic Two or More Races	1.006	0.954	1.061
MotherRaceHisp_6 Hispanic vs Non-Hispanic White	0.987	0.967	1.007
MotherRaceHisp_6 Non-Hispanic AIAN & NHOPI vs Non-Hispanic Asian	1.020	0.940	1.107
MotherRaceHisp_6 Non-Hispanic AIAN & NHOPI vs Non-Hispanic Black	0.723	0.669	0.780
MotherRaceHisp_6 Non-Hispanic AIAN & NHOPI vs Non-Hispanic Two or More Races	0.986	0.900	1.079
MotherRaceHisp_6 Non-Hispanic AIAN & NHOPI vs Non-Hispanic White	0.967	0.896	1.043
MotherRaceHisp_6 Non-Hispanic Asian vs Non-Hispanic Black	0.708	0.683	0.735
MotherRaceHisp_6 Non-Hispanic Asian vs Non-Hispanic Two or More Races	0.966	0.910	1.026
MotherRaceHisp_6 Non-Hispanic Asian vs Non-Hispanic White	0.947	0.917	0.979
MotherRaceHisp_6 Non-Hispanic Black vs Non-Hispanic Two or More Races	1.364	1.292	1.439
MotherRaceHisp_6 Non-Hispanic Black vs Non-Hispanic White	1.337	1.309	1.366
MotherRaceHisp_6 Non-Hispanic Two or More Races vs Non-Hispanic White	0.981	0.931	1.033
TermOfFirstPrenatalVisit 3rd Trimester Visit vs First Trimester Visit at NumOfPrenatalVisits=1-6 Visits	0.234	0.223	0.246
TermOfFirstPrenatalVisit 3rd Trimester Visit vs Second Trimester Visit at NumOfPrenatalVisits=1-6 Visits	0.369	0.351	0.389
TermOfFirstPrenatalVisit First Trimester Visit vs Second Trimester Visit at NumOfPrenatalVisits=1-6 Visits	1.575	1.511	1.643
TermOfFirstPrenatalVisit 3rd Trimester Visit vs First Trimester Visit at NumOfPrenatalVisits=11-16 Visits	1.256	1.067	1.478
TermOfFirstPrenatalVisit 3rd Trimester Visit vs Second Trimester Visit at NumOfPrenatalVisits=11-16 Visits	0.970	0.822	1.145
TermOfFirstPrenatalVisit First Trimester Visit vs Second Trimester Visit at NumOfPrenatalVisits=11-16 Visits	0.773	0.745	0.801
TermOfFirstPrenatalVisit 3rd Trimester Visit vs First Trimester Visit at NumOfPrenatalVisits=17+ Visits	1.179	0.705	1.973

	0.751	0.444	1.272
TermOfFirstPrenatalVisit First Trimester Visit vs Second Trimester Visit at NumOfPrenatalVisits=17+ Visits	0.637	0.564	0.720
TermOfFirstPrenatalVisit 3rd Trimester Visit vs First Trimester Visit at NumOfPrenatalVisits=7-8 Visits	0.210	0.191	0.231
TermOfFirstPrenatalVisit 3rd Trimester Visit vs Second Trimester Visit at NumOfPrenatalVisits=7-8 Visits	0.402	0.364	0.443
TermOfFirstPrenatalVisit First Trimester Visit vs Second Trimester Visit at NumOfPrenatalVisits=7-8 Visits	1.912	1.835	1.992
TermOfFirstPrenatalVisit 3rd Trimester Visit vs First Trimester Visit at NumOfPrenatalVisits=9-10 Visits	0.439	0.385	0.500
TermOfFirstPrenatalVisit 3rd Trimester Visit vs Second Trimester Visit at NumOfPrenatalVisits=9-10 Visits	0.607	0.531	0.694
TermOfFirstPrenatalVisit First Trimester Visit vs Second Trimester Visit at NumOfPrenatalVisits=9-10 Visits	1.383	1.333	1.435



Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG

0.500	2003	836E3	1602	80677	91.1	2.4	99.8	44.4	8.8

Fit Statistics for SCORE Data

Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
WORK.VALIDATIONDATASET	393904	-107632	0.0891	215351.4	215351.4	215830.3	215830.3	0.054936	0.121318	0.715568	0.075435
WORK.TESTDATASET	920341	-252145	0.0894	504377.2	504377.2	504893.4	504893.4	0.05482	0.120869	0.716561	0.075732

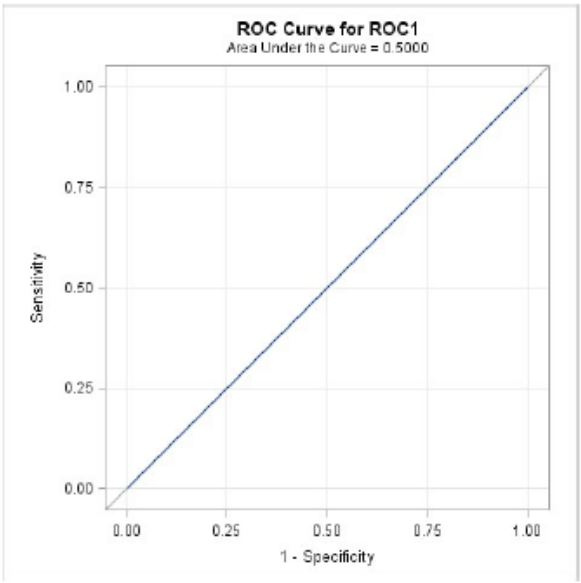
ROC Model: ROC1

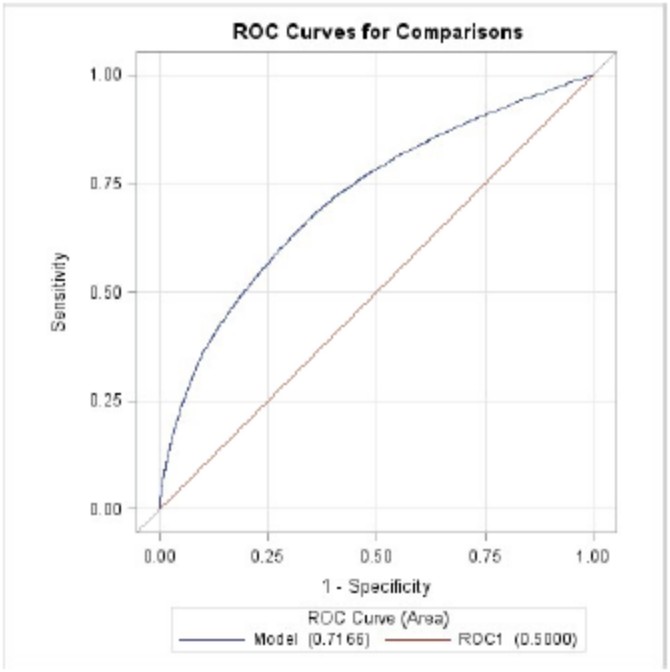
Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

-2 Log L	=	556178.34

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.3156	0.00365	403515.745	<.0001





ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D (Gini)	Gamma	Tau-a
Model	0.7166	0.000980	0.7146	0.7185	0.4331	0.4339	0.0708
ROC1	0.5000	0	0.5000	0.5000	0	.	0

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Reference = Model	1	48856.2105	<.0001